DOCUMENT RESUME

ED 212 659 TM 820 054

AUTHOR	Winters, Lynn
TITLE	The Effects of Differing Response Criteria on the Assessment of Writing Competence.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
PUB DATE	Nov 78
GRANT	OB-NIE-G-78-0213
NOTE	127p.
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	*Comparative Analysis; Higher Education; High Schools; Holistic Evaluation; *Scoring; Student Placement; *Test Validity; Writing (Composition); *Writing Evaluation
IDENTIFIERS	CSE Analytic Scale; Diederich Rating Scale for Essays; General Impression Scoring; Inter Rater Reliability; T Units

ABSTRACT

The purpose of this study was to investigate the relative validities of four essay scoring systems, reflecting alternative conceptualizations of the writing process, for identifying "competent" writers. Each rater was trained in two of the four scoring systems: General Impression Scoring (GI), Diederich Expository Scale (DES), CSE Analytic Scale (CSE), or T-Unit Analysis (T). Writing samples were randomly selected from four composition classes: low performance high school writers, high performance high school writers, low performance college writers, and high performance college writers. The experimental design was a hierarchical mixed model: all subjects were crossed with all raters, while topics and scoring systems were nested within raters. All four scoring systems had high reliability. Different scoring systems produced different patterns of results, which may mean that the writing performance of each group cannot be adequately described by the same scoring system. Three of the systems (GI, DES, and CSE) were associated with some differences in group performance; T-Unit was not, making it virtually useless for classification purposes. The classification accuracy of the three systems (GI, DES, CSE) was questionable in distinguishing all but the high school low group. (BW)

****
*   Reproductions supplied by EDRS are the best that can be made   *
*   from the original document.   *
****

ED212659

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

DELIVERABLE: November 1978

STUDIES IN MEASUREMENT AND METHODOLOGY

Work Unit 1:
Design and Use of Tests

Eva Baker and Edys Quellmalz
Project Directors

THE EFFECTS OF DIFFERING RESPONSE CRITERIA ON THE ASSESSMENT OF WRITING COMPETENCE

Lynn Winters

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. S. Gray

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Grant Number:
OB-NIE-G-78-0213

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

TM 820 014

The work upon which this publication is based was performed pursuant to a grant with the National Institute of Education, Department of Health, Education and Welfare. Points of view or opinions stated to not necessarily represent official NIE position or policy.

TABLE OF CONTENTS

Introduction 1

Related Literature 4

  Section 1. The Assessment of Writing Skill 4

    The Reliability of Writing Samples 5

      Writer Consistency 7
      Assignment Effects 8
      Rater Consistency 9
      Scoring Systems 11

  Section 2. The Validation of Scoring Systems for Assessing Writing Skill 16

    The Use of Criterion Groups to Validate Scoring Systems . . . 20

Design and Procedures 22

  Research Questions 22

  Research Design 22

    Variables 23
    Controls 24
    Design 30
    Subjects 30
    Writing Tasks 31

  Dependent Measures 32

    General Impression Scoring 32
    Diederich Expository Scale 34
    CSE Analytic Scale 36
    T-Unit Analysis 37

Analyses 39

  Rater Reliability 39

  Scale Intercorrelations 40

  The Performance of Criterion Groups Across Scoring Systems . . . 42

  The Efficiency of Scoring Systems for Predicting Criterion Group Status 44

    Multivariate Analysis of Variance 44
    Discriminant Classification 46

  Summary of Results 47

Conclusions 69
  Reliability 69
    The Effect of Differing Response Criteria on Patterns of Writing Performance 71
  Validity 73
    Classification 74
  Recommendations 78
  Suggestions for Further Research 81

References

Appendices

A. Directions to Students and Writing Prompts

B. Agenda for Rater Training

C. Dependent Measures
   1. Impressionistic Rating Scale and Amended Rubric
   2. Educational Testing Service Composition Scale (Diederich Expository Scale)
   3. Center for the Study of Evaluation Analytic Scale
   4. T-Unit Analysis Rating Procedures

D. Results of Pilot Tests for Rater Reliability

FIGURES

1. Raters Assigned to Training Sessions (D)
2. Assignment of Writing Samples to Raters (D)
3. General Impression: Comparison of Score Distributions Across Criterion Groups
4. Diederich Expository Scale: Comparison of Score Distributions Across Criterion Groups
5. CSE Analytic Scale: Comparison of Score Distributions Across Criterion Groups
6. T-Unit: Comparison of Score Distributions Across Criterion Groups
7. High School Low: Comparison of Z score Distributions Across Scoring Systems
8. High School High: Comparison of Z score Distributions Across Scoring Systems
9. College Low: Comparison of Z Score Distributions Across Scoring Systems
10. College High: Comparison of Z Score Distributions Across Scoring Systems

TABLES

1. Distribution of Papers to Rater Pairs
2. Data for Estimating Rater Reliabilities of Scoring Systems
3. Intercorrelations Among Scoring Systems: Total Sample
4. Intercorrelations Among Scoring Systems: High School Low
5. Intercorrelations Among Scoring Systems: High School High
6. Intercorrelations Among Scoring Systems: College Low
7. Intercorrelations Among Scoring Systems: College High
8. Criterion Group Performance Across Scoring Systems Means and Standard Deviations
9. Analysis of Variance: System 1 GI
10. Analysis of Variance: System 2 DES
11. Analysis of Variance: System 3 CSE
12. Analysis of Variance: System 4 T
13. F Ratio for Multivariate Tests of the Equality of Mean Vectors
14. Results of Discriminant Analysis Within Groups Correlation Matrix
15. Results of Discriminant Analysis, LDF's System 1-3
16. Results of Discriminant Analysis, LDF's System 2-3
17. Classification Accuracy for GI, DES, CSE Scoring Systems
18. Classification Accuracy for DES, CSE Scoring Systems

THE EFFECTS OF DIFFERING RESPONSE CRITERIA ON THE ASSESSMENT OF WRITING COMPETENCE

INTRODUCTION

The assessment of writing skill is a task more akin to slaying the Hydra or cleaning the Augeuan stables than a straightforward application of modern measurement technology to the problem of testing composition skill. Few other areas of the school curriculum remain as impervious to reliable and valid measurement as writing competence. The problems of judging the adequacy of written composition begin with the difficulty in accounting for the many sources of variation influencing writing and are compounded by the numerous competing definitions of "writing skill." Writing ability is a complex and ill-defined construct. Associated with the attempt to describe writing skill are such issues as, what are the features of skilled writing? What is involved in the act of writing? (Braddock, Lloyd-Jones, and Schoek, 1963). What does it mean to write a piece of discourse? Where is the point of mastery in writing? (Cooper, 1975).

In attempting to answer these questions, researchers have adopted diverse strategies derived from alternative conceptualizations of the writing process. Cooper (1977) conceives of writing as a holitistic process whereby, "a piece of writing communicates a whole message with a particular tone to a known audience for some purpose (p. 3)." To holistic researchers, the evaluation of writing skill cannot be divorced from human judgment about the total effectiveness of a piece of work. In contrast to this view is the analytic approach which posits that a good written product possesses certain elements existing apart from the reader's reaction to the whole writing sample. Lloyd-Jones, who exemplifies this approach, defines writing skill as "discoursing (1977)." For Lloyd-Jones, the characteristics of good writing vary with the purpose of the message; writing is to be judged as to how well it incorporates the characteristics thought to belong to a given rhetorical task. In a third conceptualization of the writing process, the linguistic approach, the good writer is one who produces "syntactically mature" sentences. Writing is evaluated in terms of the average length of independent clauses, with the more "mature" writers producing longer syntactic units (Hunt, 1977).

These competing approaches to the evaluation of writing raise several issues related to the validity of decisions made using a particular method. Among these are: Do decisions about writing competence depend upon the scoring procedure selected? Are scoring systems valid in and of themselves, or do they possess different validities, depending upon the writing task or writer population about which the decision is to be made? Is there a single scoring system appropriate for several decision purposes?

If scoring systems derived from alternative conceptualizations of the writing process do possess different validities, then several instructional issues are raised: Should different writing goals be defined for different groups of students? Which instructional strategies are most appropriate for meeting these goals? Which scoring systems provide the most pertinent information for instructional decision-making? Which writing tasks are most important for different writing assignments? Credible writing assessment programs, then, must be concerned not only with the selection of valid and reliable instruments but also with the relationship of these measures to the kinds of instructional decisions that will be made from test results.

Finally, the selection of appropriate scoring strategies for assessing writing skill is dependent upon how "writing competence" is to be defined. There are two general strategies for defining writing competence, judgmental and empirical. Traditionally, the schools have relied on "expert" judgment, in the form of teacher grades, to identify the competent writer. However, with the installation of competency-based testing programs, there is an implication that competence is something more than one's ability to reach a pre-specified standard of excellence (Linn, 1977). There is the underlying assumption that the competent student is one who is able to perform the target skill effectively after leaving school. The transfer of student writing performance to post high school tasks has not been investigated. It is, nevertheless, an issue which begs exploration, given the current demand for information about student writing "competence" in state-mandated testing programs and the widespread practice of placing students in university composition programs on the basis of writing samples. Furthermore, the comparison of the writing performance of several different levels of writers may provide test makers with some guidelines upon which to base judgments about what constitutes defensible definitions of writing competence.

The purpose of this study was to investigate the relative validities of several scoring systems, reflecting alternative conceptualizations of the writing process, for identifying "competent" writers. Groups of high school and college writers identified a priori as representative of different competence levels were the criterion against which the systems were validated. The accuracy of the scoring systems for describing the different kinds of writing performance, for representing/discriminating among them, and for identifying writers belonging to the criterion groups were investigated by the following research questions:

1) What is the pattern of performance of a priori identified groups of writers across different scoring systems?
2) What is the pattern of performance of these groups on the same scoring system?
3) Which scoring systems discriminate most clearly among criterion groups?
4) Which scoring systems are the best predictors of criterion group status?

RELATED LITERATURE

The literature related to the validation of essay scoring systems is reviewed in two sections. The first section presents current strategies for measuring writing skill with an emphasis on three different approaches to scoring writing samples. Section two focuses on procedures for investigating the empirical validities of the scoring criteria discussed in section one.

Section 1.

The Assessment of Writing Skill

Debate persists about whether objective or essay tests provide the more appropriate and efficient measure of writing skill. Objective tests as a measure of the construct "writing ability" are under attack as the sole source of information for determining whether students are competent writers. The position of those who favor essay tests was summarized by Paul Diederich:

> As a test of writing ability, no test is as convincing to teachers of English, to teachers in other departments, to prospective employers, and to the public as actual samples of each student's writing (p. 1, 1974).

This position is further strengthened by the results of a study conducted by Richard Braddock to review selected, published English composition tests. Included in the review is an unabashed indictment of objective writing tests:

> No commercially available standardized test attempts to measure a student's ability to organize and detail his or her writing so that prospective readers can share the writer's experience and appreciate his or her purposes. At this stage of our understanding of writing and of testing, it is difficult to believe that any standardized (i.e. objective) test will be constructed which can measure such abilities...therefore, anyone who professes to measure writing ability with a standardized test is either telling a falsehood or speaking from ignorance. It

> is a cruel joke that some corporations are today selling school boards on the notion they can do just that. And it is ironic that some colleges are excusing students from composition requirements merely on the basis of standardized test scores (p. 118, 1976).

The Reliability of Writing Samples

The decision to substitute essay for objective tests to assess writing skill, while satisfying requirements of construct validity, poses problems for reliable measurement. The use of writing samples introduces many sources of measurement error that affect the interpretability of test scores. More importantly, the validity of decisions based on unreliable scores is impugned.

Measurement error associated with the rating of essays has been described in recent summaries of writing research (Cooper, 1975; French, 1972; Coffman, 1971; Braddock, Lloyd-Jones, & Schoer, 1963). Cooper (1975) listed several variables affecting the stability and accuracy of essay scores. Among these are the mode of discourse of the assignment, the amount of writing time allowed, the tendency of writing performance to be inconsistent, and the expectations and background of the raters.

French (1972) noted the presence of the same intervening variables but describes them as 1) student error, i.e., writer inconsistency, 2) reader disagreement, 3) test error, that is, performance variation attributable to the writing assignment, and 4) scale error, the severity of judgment of individual readers.

The classification scheme of Braddock et al. is analogous to French's, but includes two additional sources of variation. Their classification lists 1) the writer variable, 2) the assignment variable, including the level of abstraction of the topic, topic familiarity, the mode of discourse, and the time available for writing, 3) the rater variable, defined as rater consistency and 4) the colleague variable, or between rater agreement.

Coffman's categories focused primarily on the threats to reliability posed by the raters alone. He described three tendencies common to raters that threaten the stability of writing scores: 1) Different raters tend to assign different grades to the same paper on different occasions. 2) A single rater tends to assign different grades to the same paper on different occasions. 3) The differences in the assignment of grades tend to increase as the essay question permits greater freedom of response (p. 277, 1971). Coffman also developed more fully than Cooper, Braddock, or French the reasons for inter-rater disagreement. He described how differences among the raters grading the same writing sample occur as a result of differences in their own internalized standards of excellence (severity). Another internal rater variable relates to the way in which grades are distributed. Some raters assign grades closely around their average mark, while others spread scores more widely. A final source of rater variation is the disagreement that occurs when scoring criteria are not explicit, resulting in the same paper receiving several different marks.

To summarize, both French and Braddock identified four major sources of variability affecting the judgment of writing samples. However, two of these are associated with the rater alone. Coffman dealt with the reliability issue almost exclusively in terms of rater variation. This emphasis on rater variability and lack of attention to other error sources derived from the fact that most writing research is concerned with the various types of rater inconsistency rather than variables such as the writer, assignment and testing occasion. While it is true that some studies (Kincaid, 1963; Troxler, 1935) have discussed writer and assignment effects on essay scores, these variables are of secondary importance when their contribution to test error is compared to that of rater variation. For this reason, the reliability estimate of interest in the present study was rater agreement. The following sections briefly review the literature on writer and assignment effects because these variables must be controlled or accounted for in any writing study. However, the focus of the literature review on reliability is on how differences among raters can be minimized in order to increase the generalizability of essay scores.

Writer Consistency

The writer variable is defined as the fluctuation in the quality of an individual's writing performance between testing occasions. In a study to investigate the validity of the use of a single topic to assess writing performance, Kincaid (1963) also examined the differences in scores between two papers written by the same student when both topic and rater were controlled. It was found that the performance of students did vary significantly between writing occasions, especially the performance of the better writers. Kincaid concluded that a single paper written by a student on a given topic at a particular time could not be considered a valid basis for evaluating writing achievement. Anderson (1960) analyzed the writing performance of 55 eighth grade students on eight occasions and found that 71% of the sample showed evidence of composition fluctuation. These results supported Kincaid's findings.

While studies such as those conducted by Kincaid and Anderson point to the existence of day to day variability in writer performance, and the need to have more than one sample, there are no writing sample studies that isolate the optimal minimal number of writing papers necessary to obtain reliable estimates of an individual's writing skill. Troxler (1935) suggested that as many as 40 compositions were required to estimate the general ability level of a writer based on obtained reliability estimates of .60 for one piece of writing. Unfortunately, taking many samples is impractical for assessing writing skill in most research settings; therefore, the control for writer inconsistency most commonly employed is the guideline provided by Cooper (1977), Braddock, et al.,1963) and Diederich (1974): At least two writing samples, written on different days and given at least two independent ratings, should serve as the minimal number of samples and ratings for estimating writing performance. This rule of thumb assures a modicum of control for writer variation until additional evidence about required number of writing samples is obtained.

It should be added that two other factors, writing speed and verbal fluency (Coffman, 1971), have been suggested as having possible effects on the stability of writing performance. Unfortunately, little research has been conducted in these areas.

Assignment Effects

At least three factors associated with the writing assignment influence writing performance, 1) the type of topic, 2) the mode of discourse, and 3) the amount of time allocated for writing (Braddock,et al., 1963). Wiseman and Wrigley (1958) reported significant variations in the writing performance of 11 year olds who wrote on different topics. While they concluded that differences in writer performance associated with the choice of different topics were attributable to "writer ability" rather than topic difficulty, the alternative hypothesis is not ruled out. In the absence of evidence that topic differences among students and between testing occasions do not influence writing performance, safeguards such as the selection of topics "familiar" to the writer population, and the assignment of the same topic to all examinees appear to be necessary when doing writing research.

A second aspect of the assignment that affects writing performance is the mode of discourse of the assigned topic (e.g., argumentation, exposition, narration, description). There is evidence that the mode of discourse has at least two effects on writing performance: 1) The sentence structure is affected by the mode of discourse (Frogner, 1933; Seegers, 1933; Hunt, 1964), 2) some modes of discourse are more "difficult" than others (Veal& Tillman, 1971).

The assumption that writing skill may be discourse-specific is reflected in the popularity of such discourse related scoring systems as the Diederich Expository Scale and Primary Trait Scoring. The high interrater reliabilities obtained when these systems were used (Lloyd-Jones, 1977; Diederich, 1974) suggest that some of the unreliability reported in relation to the assignment can be controlled by the use of assignment-specific scoring criteria. Current practice in writing research therefore, controls for topic and mode of discourse (O'Hare, 1973).

Rater Consistency

The source of variation receiving the most attention in research on the reliability of scoring essays, and of major concern to this study, is that associated with the rater. The rater variable has two major components 1) within rater inconsistencies and 2) between rater disagreement (Findlayson, 1951). Of these two aspects, between rater disagreement has been the most frequent subject of research.

Two widely cited studies, one that demonstrates the tendency of different raters to assign different marks to the same paper and another that provides a model for controlling rater disagreement, were those conducted by Diederich (1961) and Follman and Anderson (1967). Diederich instructed 53 raters, representing six different professions, to sort 300 essays (150 on each of two topics) into nine grading categories. He found that no one paper received less than five different grades and that 94% of the papers received seven or more of the nine possible grades. Inter-rater reliability was reported as .31, which represents the median correlation between readers.

While the Diederich findings might suggest that essay grading is beset with problems in obtaining rater agreement, the Follman and Anderson study provides a more hopeful conclusion. In a study to investigate the effects of theme evaluation procedures on the reliability of ratings, Follman and Anderson assigned 25 raters, in groups of five, to one of five rating systems. Each rater graded 10 themes. It was found that the essays received similar scores regardless of the rating system used, and that there was high within group rater consistency among the five raters in any one group (interclass correlations from .81 to .95). The conclusion reported was that the variability observed in the Diederich study may not have been a result of the systematic tendency of two people to evaluate the same writing sample differently. Follman and Anderson posited that it was the heterogeneous background of the raters (who had been selected because they did represent different fields) that was responsible for the unreliable essay ratings. The researchers then suggested that the use of systematic theme evaluation procedures (i.e., explicit scoring criteria) might control rater variation because it would serve "to sensitize and organize the rater's perceptions and give direction to his attitude and judgment..." (Follman & Anderson, p. 199, 1967). Subsequent studies incorporating the recommendations of Follman and Anderson have reported high correlations between raters judging the same piece of writing (Conlan, 1976).

Scoring Systems

The discussion of sources of variation associated with the reliable assessment of writing competence in the previous section includes the recommendation of Follman and Anderson that the use of a systematic scoring procedure can increase the reliability of rater judgments. The fact that training of raters is an effective way to improve rater agreement is underscored by Cooper and Odell (1977):

> As emphatically as I can, then, let me correct the record about the reliability of holistic judgments. When raters are from similar backgrounds and when they are trained with a holistic scoring guide...they can achieve scoring reliabilities in the high eighties and low nineties on their summed scores from multiple pieces of a student's writing (p.19).

However, the use of scoring criteria is important not only for training raters to make stable estimates of writing skill, but to provide a better description of the writing skill being tested. Particular scoring systems represent different approaches to the identification of the salient characteristics of good writing. Decisions about writing competence derived from one scoring system may or may not lead one to the same decisions about an examinee when another system is used. What is still at issue is how dependent are decisions about an examinee's writing skill upon the particular scoring system used? The following discussion of three approaches to the scoring of essays will highlight the differences among these approaches in order to better illustrate this issue.

Scoring systems are generally of three types, impressionistic, analytic, or frequency counts (Braddock, et al., 1963). General impression marking is the system whereby two or more raters quickly read and rate an essay, then assign a single score ranking the writing sample in relation to all other papers being evaluated. The several ratings assigned by different scorers are averaged to obtain a reliable (i.e. stable) estimate of a writer's ability.

The number of rating categories used in general impression marking varies considerably. Four point rating scales have been used in research (Meyers, McCanville & Coffman, 1966) and for the English Composition Test (Conlan, 1976). Both NAEP and the Advanced Placement Test in English use an eight point scale. Regardless of the point spread, the one feature shared by impressionistic systems is the inclusion of rater training prior to scoring. This training consists of the study and discussion of sample essays by the readers in order that they become "calibrated" to reach consensus on score assignment.

Several sorts of general impression marking are popular, some with rubrics for raters, others without. The most famous of the unguided general impression systems is the Everyman's Scale used in the Follman and Anderson study (1967) cited earlier. Raters were given the instruction to judge essays by whatever criteria they chose, with the provision that grades be assigned on a scale from 1 to 5 and that at least one paper be put into each of the five categories. Using a more structured approach, Godshalk, Swineford, and Coffman (1966) asked judges to rate papers on a scale of 1-3 with copies of sample essays exemplifying each rating category serving as the rubric. The more formal version of the Godshalk approach is the California Essay Scale (Nail et al., 1960). Raters are provided with model essays representing six levels of writing quality. Their task is to match examinee samples to the model essays in order to determine the appropriate score. As in the systems described earlier, extensive training precedes actual scoring in order to obtain high rater agreement.

Analytic scoring systems, in contrast to general impression marking, require careful scrutiny of the writing samples. While general impression systems are predicated on the assumption that good writing is clearly recognizable, analytic systems assume that quality writing is characterized by the inclusion of certain rhetorical elements. The elements of interest are specified on the rating scale. Points are assigned by readers according to the observed presence or absence of the elements. The advantage of using an analytic scale over impressionistic scoring is that it can provide diagnostic information to students for improvement of writing skills. Two studies (Alpren, 1973; Sager, 1973) reported that the use of teacher or peer feedback from analytic scales was helpful to students in improving their writing.

One of the most popular discourse-specific scoring systems is the ETS composition scale, often called The Diederich Expository Scale. The scale emerged from a study that investigated factors influencing human judgment of expository writing samples. Sixty readers, representing the fields of college English, social science and natural science, business, law, and editing, were asked to rate 300 papers using only their own judgment. Raters were also instructed to write any comments they wished on the papers. A factor analysis of patterns of rater agreement, combined with a review of the comments on the essays by raters belonging to each of the five obtained clusters, was used to form the categories for the Diederich Expository Scale (Diederich, 1974).

Another discourse related analytic scale is Primary Trait Scoring, developed by the National Assessment of Educational Progress to score large numbers of writing samples. Primary Trait protocols are generated by selecting a discourse model and identifying the characteristics of the assigned writing task as they relate to the target mode of discourse. The scoring guide defines and describes the levels of quality for the writing task. Raters are asked to indicate what level of quality a writing sample exhibits for each rhetorical task described by the Primary Trait protocol (Lloyd-Jones, 1977).

The third type of scoring system is the frequency count. Whereas general impression and analytic scales focus on the quality of a writing sample, frequency counts are concerned with the quantity of certain linguistic elements. Such units as the number of unique words in an essay, the number of sentences, and the number of words per clause (T-units) become measures of writing skill. Although at first glance it may not appear that the number of linguistic elements present in a sample is related to writing skill, research does not bear this out. (Howerton, et al., 1977)conducted a study to investigate the relationship between quantitative and qualitative measures of writing skill. One thousand subjects selected by a stratified random sampling process to match national norms were asked to write a 200-250 word essay using the "famous person" stimulus from the 1972 NAEP writing study. The essays were scored in several ways. Impressionistic scoring was done using the ETS Composition Evaluation Scale. An analytic system, the Diederich Expository Scale, provided a second measure of writing quality. Quantitative measures included counting the total number of words per writing sample, the total number of sentences, the number of unique words per composition, and the mean T-unit length. Quantitative score units were then used as predictors of qualitative scores. Stepwise regression procedures were employed to extract the best set of quantitative predictors. It was found that several quantitative and qualitative measures of writing skill were significantly related. The best set of quantitative predictors were the total number of words in the sample and the number of words per T-unit. These results support research conducted by O'Hare (1973), Hunt (1964, 1965) and Combs (1976) which indicate that positive relationships exist between such syntactical units as the T-unit and writing quality as measured by holistic or analytic systems.

The frequency counting method with perhaps the most potential for judging levels of writing skill is T-unit analysis. The T-unit is a single main clause plus whatever subordinate clauses or nonclauses are attached to or embedded within that one main clause (Hunt, 1977). The T stands for terminable. T-units are the shortest grammatically complete sentences that a passage can be cut into without creating fragments. The T-unit length hypothesis is that "as school children get older, the T-units they write tend to get longer, measuring length as the mean number of words per T-unit (Hunt, p. 93, 1977)." Studies investigating the relationship of T-unit length to language development and to writing quality have tended to support the hypothesis that they are positively related to T-unit length (Bateman & Zidonis, 1964; Mellon, 1965; O'Donnell, Griffin & Novis, 1967, O'Hare, 1970).

A number of studies (Follman & Anderson, 1967; O'Hare, 1973; Howerton, Jackson, & Selden, 1977) have found relationships among the three types of scoring systems.

These findings have led researchers to hypothesize that different scoring systems measure a number of elements in common (Follman & Anderson, 1967). On the other hand, the Diederich Expository Scale was not highly correlated with the impressionistic systems used in the Follman study (California Essay Scale, Cleveland Rating Scale, Follman English Mechanics Guide and Everyman's Scale). Significant correlations between mean T-unit length (especially for younger writers) and impressionistic ratings on the Composition Essay Scale (ETS) were found in the Howerton study. However, a simultaneous comparison of impressionistic, analytic, and frequency count scoring systems has not been done. The different assumptions about writing skill implicit in each type of system suggests that high intercorrelations may not exist, especially when the comparison among them is direct. One subissue of interest in this study is, to what extent are the three types of scoring systems measuring different aspects of writing skill?

Section 2.

The Validation of Scoring Systems for Assessing Writing Skill

Scoring systems may possess several sorts of validities, depending upon the kinds of interpretations they are expected to yield. There are at least five important decisions for which writing samples are used:

1) Evaluation of the effectiveness of alternative instructional programs
2) Screening of students for college entrance and/or placement in English composition programs
3) Identification of students as competent or not competent for graduation or promotion

4) Description of the typical writing performance of a large national sample

5) Identification of students needing remedial instruction.

Not all of these decisions are well-served by the same measures. For the first two purposes, inferences need only be made about the adequacy of group performance. Purposes two and three are sometimes met by techniques that rank student papers, rather than rate them against explicit criteria (Conlan, 1976; Smith, R., 1975). Identification of skill deficiencies requires that the measure yield descriptive information about patterns of student strengths and weaknesses. Certification of competence for graduation or promotion suggests that some kinds of predictions about the adequacy subsequent writing performance be made.

There have been few attempts to validate direct measures of writing skill (i.e. sets of response criteria) beyond the level of content or descriptive validity. For example, the content validity of the response criteria used in NAEP's Primary Trait Scoring was established by "expert judges" (Lloyd-Jones, 1977), as were those used by the College Entrance Examination Board for the Composition test and for the Advanced Placement English test given by ETS. T-unit analysis has been validated against the Diederich Scale or some system of impressionistic scoring in several studies (Hunt, 1970; Howerton, et al., 1977). However, none of these well-established scoring systems has been investigated as to its ability to predict success on subsequent writing tasks, a concern of current graduation competency testing programs. Thus, there is a dearth of information related to the empirical validities of essay scoring criteria.

While few guidelines exist in the literature on writing research for the empirical validation of scoring systems, techniques for establishing empirical validity have been described in the context of criterion-referenced testing (Hambleton,et al., 1975; Harris, 1974; Millman, 1974; Popham, 1975, Wilson, 1974). Criterion-referenced measurement technology offers two major strageties for investigating empirical validity: 1) The use of scores on established tests as a criterion to which newly-developed test items should predict and 2) the use of the performance data of criterion groups to validate the accuracy of classification decisions to be made from a test.

The use of test scores as a criterion for empirical validation is adapted from the field of industrial psychology. Traditionally, classification decisions in industrial psychology have been made by examining the joint distribution of criterion outcomes and test scores to determine what decision rule produces the most satisfactory distribution of outcomes (Cronbach, 1971). The criterion outcomes may be selected in any one of several ways. Cronbach identifies a procedure called "empirical keying." In empirical keying, a person or group having the criterion characteristics, such as high supervisor ratings for job performance, are administered a set of test items. The responses of this group are then analyzed to determine which correlate most highly with criterion group membership. Those items with high correlations are then selected to become part of a test used to select people most likely to possess the attributes of the criterion group.

A similar procedure has been described by Perry et al., (1968). Examinee groups with contrasting characteristics on the criterion variable are chosen. The test being validated is administered to both groups, then responses of the contrasting groups are compared in order to select those items which predict to the appropriate groups.

The use of criterion groups for construct validation of tests in education has been suggested by several authors (Harris, 1974; Linn, 1977; Millman, 1974). The criterion groups of interest in education have generally been those groups of students representing different levels of instruction on the task of interest. The assumption underlying the selection of criterion groups is that uninstructed examinees may be expected to have less of the skill, while instructed students will have more. It is predicted that the difference in the amount of the skill in question will be reflect in different test scores on the unobservable variable (Millman, 1974).

The strategy of using uninstructed versus instructed groups of examinees as a criterion for establishing empirical validity is not without problems. Millman (1974) identified three shortcomings of this method. The criterion group method assumes that examinees in the instructed group are indeed the products of effective instruction. Poor performance of the instructed examinees may mean that the test is a poor measure of the skill or that the examinees did not represent a truly "instructed" group. Secondly, measures whose empirical validities are tied to criterion groups have instructionally dependent interpretations. That is, they are meaningless when used with examinees who have not received instruction in the target skill. However, when the intent of a test is to identify students in order to provide appropriate instructional treatments, these shortcomings become a strength.

Still another weakness of using identified instructed versus uninstructed learners for empirical validation is that for most skills assessed at the secondary level, all students have received some instruction. The selection of examinees who possess little or none of the construct of interest can be most difficult.

A solution that meets many of the above challenges to the use of pre-post instructed examinees has been identified by NAEP. Criterion groups were selected to represent examinees at different grade levels and ages. The rationale for this procedure is summarized by Wilson (1974):

> Assumptions behind the selection of...(examinees of different ages) is that older examinees have greater training in a given skill than younger examinees, when training in the target skill, such as reading, writing, or arithmetic, is a continuous process. Thus, response data obtained is expected to yield not dichotomous patterns, as between instructed vs. uninstructed, but an increasing percentage of correct responses from the age groups selected (p. 36.)

The Use of Criterion Groups to Validate Scoring Systems

The use of criterion groups to investigate the empirical validities of different scoring criteria for making decisions about writing competence provides an *a priori* classification system against which the scoring criteria can be tested. Conversely, the scoring systems describe competing interpretations of the construct "writing competence," which, when linked with the performance of different writing groups, may help explain the differences found among different levels and kinds of writers. Empirical validation, then, provides a method whereby the consequences of using different essay scoring criteria may be examined. The growing practice of using writing samples in large scale and graduation competency testing programs creates a need for documentation of the consequences of using a particular set of scoring criteria. This documentation allows decision makers to select the scoring system most appropriate to the decision purposes of the testing program.

The four scoring systems validated in the present study were selected to represent the three types of scoring criteria: impressionistic, analytic, and frequency counting. Two types of analytic scales, the Diederich Expository Scale and the CSE Analytic Scale, were tested. The Diederich Scale contains both impressionistic elements and elements tied to the requirements of "exposition." The CSE Analytic Scale was derived from current textbooks on teaching written composition and emphasizes the elements of writing currently taught. T-unit analysis was the method selected to exemplify the frequency counting approach. It was selected because it has been validated against qualitative measures of writing and was expected to have elements in common with impressionistic scales and the Diederich scales. Additionally, T-unit analysis has proved to have important implications for improving writing skill.

The criterion against which scoring systems were validated is their ability to sort a priori identified examinees, representing different levels of writing skill, into the correct writing groups. Criterion groups in the study were chosen to contain low and high performance high school and college writers. These four groups provided an operational definition of the construct "writing competence." Criterion group members who had successfully met high school or college demands were considered "competent." Writers in other groups whose performance matched that of competent examinees might also be judged "competent" for certain writing purposes. It was hoped that by using criterion groups, this study would clarify what is meant by the term "writing competence," and identify those scoring systems which are most accurate in discriminating among different levels of writing skill and classifying examinees into the appropriate compentence categories.

DESIGN AND PROCEDURES

Research Questions

This study investigated the validities of four essay scoring systems for making decisions about writing skill. The performance of criterion groups, a priori identified groups of writers with different levels of writing skill, represented the criterion against which the systems were validated. Scoring systems were tested for their ability to classify members accurately into the pre-identified groups. A second concern of the study was the identification of those systems which were able to discriminate among different levels of writing skill as defined by criterion groups. Four questions were investigated:

1. What is the pattern of performance of each criterion group across the different scoring systems?
2. What is the pattern of performance of different criterion groups on the same scoring system?
3. Which scoring systems discriminate most clearly among criterion groups?
4. Which scoring systems are the best predictors of criterion group status?

Research Design

The study was designed to examine the writing performance of four groups of writers on four different scoring systems. Writing samples were obtained on two parallel expository topics taken on two testing occasions, one week apart. Two sets of four raters were assigned to one of two training conditions, General Impression - Diederich Expository Scale or CSE Analytic-T-Unit scale, and taught to use two scoring procedures. The trained raters then read one writing sample from each of the subjects (n=80) using the scoring systems in which they were trained.

Variables

The research questions in this study focused on three issues: description, discrimination, and classification. The variables of interest differed according to the issue under investigation. The following outline summarized the variables of interest for each of the research questions.

I. Variables of Interest in Describing the Performance of Criterion Groups on Different Scoring Systems (Questions 1-2)
    A. Independent Variables: Scoring Systems as a Treatment
        1. General Impression Scoring (GI)
        2. Diederich Expository Scale (DES)
        3. CSE Analytic Scale (CSE)
        4. T-Unit Analysis (T)
    B. Dependent Variables: Scores Obtained by Subjects on Each System
        1. General Impression Scoring: Score Scale 1-6
        2. Diederich Expository Scale: Score Scale 5-45
        3. CSE Analytic Scale: Score Scale 6-24
        4. T-Unit Analysis: Score Scale 8-34 words per T-Unit
II. Variables of Interest in Finding the System(s) Which Best Discriminate Among Groups (Question 3.)
    A. Independent Variables: Criterion Groups
        1. High School Low-Performance Writers (HSL)
        2. High School High-Performance Writers (HSH)

3. College Low-Performance Writers (COLL-L)

4. College High-Performance Writers (COLL-H)

B. Dependent Variables: Scores Obtained by Subjects on Each System

III. Variables of Interest in Finding Scoring Systems Which Best Predict Criterion Group Status (Question 4.)

A. Independent Variables: Scoring Systems As A Treatment

B. Dependent Variables: Criterion Group Membership

Controls

The design did not attempt to control for the instructional history of the subjects or the effects of the testing occasion on writing performance. It did, however, attempt to control for the following:

1) Topic effects
2) Rater Background
3) Rater Unreliability

The study was designed so that topic effects could be investigated if necessary. It was hoped that the topics written for the study would be random examples of expository topics from the domain of writing tasks and that they would be of equivalent difficulty. In an attempt to create topic equivalence, specifications for the writing prompt were written. Three rules guided the generation of the prompt (see Appendix A): 1) Writing prompts will be syntactically equivalent. Differences in wording will occur only when reference is made to each of the essay topics, High School Proficiency or Drugs, Alcohol and Tobacco. 2) Topics will be ones about which both high school and college students can write with no prior instruction. 3) Prompts will be written to elicit expository responses, i.e., answers that explain and give reasons as opposed to those that provide opinions and arguments.

Other factors in addition to the writing prompt, however, influence topic difficulty. In spite of one's best attempts to anticipate what people are familiar enough with to write about, differences in personal experience affect how well a topic can be explicated. Additionally, writing performance, especially among the better writers, varies between writing occasions. As the research design provided for the collection of only two writing samples, and topics were randomly assigned to testing occasion (thus confounding occasion with topic), there existed a strong probability that topics would not be equivalent as hoped. In order to determine whether topics should be treated as an independent variable or be controlled, the data analysis on rater reliability for the first scoring system included investigation of topic effects. An analysis of variance on three factors influencing writing, topic, rater, and subject, showed that topic effects did indeed exist in the form of a topic by subject interaction. (See Appendix D.)

The confounding of topic with testing occasion in the research design precluded investigation of this interaction; therefore, it was decided that topic effects would be controlled in the rest of the study. This was done by treating topics as a fixed rather than random factor in the data analyses and by using scores that were averaged over both topics to represent the dependent measures.

Two techniques were used to control rater variation, the method of assignment of raters to scoring systems, and the use of training sessions to promote rater agreement. Since the writing samples represented the work of both high school and college students, raters experienced in teaching and evaluating high school and college composition were chosen. Six raters participated in the study: three high school teachers, two college teachers, and one college teacher who had experience teaching high school. In order to reduce rating time and minimize rater fatigue, it was decided to use four raters per scoring system, so that each rater would have to read only one essay per subject. To control for rater by scoring system bias, raters were trained in at least two systems. Thus, six raters had to be randomly assigned to at least two scoring systems so that four raters were trained in each system. Additionally, some of the scoring systems were reactive, i.e., General Impression Training could not be conducted after any other system had been learned and the two analytic scales could not be trained contiguously. These constraints resulted in the creation of two rater/training conditions: Training Session 1: General Impression-Diederich Expository Scale and Training Session 2: CSE Analytic and T-Unit. Rater assignment to training conditions were handled as follows:

1. The three high school teachers and one high school/college teacher formed a "pool" of raters to be randomly assigned to scoring conditions.

2. The two college teachers, one of remedial composition and the other of Freshman composition, were asked to participate in both training sessions and to learn all four systems so that all systems would have college teachers represented.

3. The high school teachers were then assigned to training sessions, two to a condition. Figure 1 displays the distribution of raters to training conditions.

Figure 1

Raters Participating in Training Conditions

Session 1: GI/DES

Rater 1: College Remedial Composition
Rater 2: College Freshman Composition
Rater 3: High School Teacher
Rater 4: High School Teacher

Session 2: CSE/T

Rater 1: College Remedial Composition
Rater 2: College Freshman Composition
Rater 5: High School Teacher
Rater 6: High School/College Teacher

Rater training sessions, described in the following section, were used to control inter-rater variability. These sessions were planned so that raters would practice grading papers until they reached an "acceptable" level of reliability, defined as an interclass correlation cofficient among raters of .80 or above. At the end of each training session, pilot tests for rater reliability were run on 24 sample papers, representing a random sample of both essays of three students from each criterion group.

A summary of the steps taken to control sources of variation due to topics and raters follows:

1. Examinee scores were averaged across topics.

2. A stratified random sampling procedure was used so that each rater received one writing sample from each of the 80 subjects of the study (see Figure 2) and an equal number of papers on each topic.

3. Papers were arranged in rater packages so that each rater received papers in random order. Additionally, rater packets were created so that there was no systematic "pairing" of raters on any one scoring system (see Table 1).

4. Writing samples were typed and assigned random identification numbers to obscure examinee name and criterion group status. To remove the influence of handwriting on essay scoring, writing samples were typed, errors intact, and photocopied for the raters.

5. To minimize reactivity, scoring systems were assigned to training conditions so that the two analytic scales were separated and general impression marking was introduced first.

6. Each training condition consisted of four raters, two high school and two college teachers. All raters learned at least two systems. Two raters

TABLE 1

DISTRIBUTION OF PAPERS TO RATER PAIRS BY RANDOM ORDERING PROCEDURES

Rater Pair	No. of Papers GI	No. of Papers DES	No. of Papers CSE	No of Papers T
12	33	14		
13	18	52		
14	34	18		
23	41	23		
24	16	53		
34	28	0		
15			29	33
16			47	7
25			44	8
26			36	33
56			4	40

Key: Rater 1 - College Teacher
2 - College Teacher
3 - High School Teacher
4 - High School Teacher
5 - High School Teacher
6 - High School/College Teacher

were trained in all four systems.

Design

The experimental design was a hierarchical mixed model. All subjects were crossed with all raters, while topics and scoring systems were practically nested within raters. Figure 1 displays this assignment of raters to scoring systems and Figure 2 the assignment of writing samples to raters.

FIGURE 2

Assignment of Writing Samples to Raters

Group Rater	HSL n=20		HSH n=20		Coll-L n=20		Coll-H n=20	
1	T1 n=10		T1 n=10		T1 n=10		T1 n=10	
		T2 n=10		T2 n=10		T2 n=10		T2 n=10
2								
3								
4								
5								
6								

Subjects

The subjects of the study were high school and college students enrolled in composition classes in June, 1978. Twenty (20) sets of writing samples were randomly selected from papers obtained from composition classes at a suburban high school and a major urban university. Classes were chosen to represent two levels of writing skill, high and low, for both high school and college populations. The students whose writing samples were chosen then became the criterion groups described below.

1. Low Performance High School Writers (HSL)

   Students enrolled in basic composition classes at the eleventh grade level who do not intend to attend a university.

2. High Performance High School Writers (HSH)

   Students enrolled in eleventh grade advance composition classes who plan to attend a university.

3. Low Performance College Writers (Coll-L)

   Students enrolled in remedial college composition at the University of California, Los Angeles by virtue of having failed a written English placement examination and/or having received low scores on a multiple choice composition test.

4. High Performance College Writers (Coll-H)

   Students enrolled in Freshman composition at the University of California, Los Angeles after having passed the placement examinations or a course in remedial composition.

Writing Tasks

Subjects were given the task of producing two, 200 word expository writing samples on two 50 minute testing occasions, one week apart. Two parallel topics, developed at CSE (Smith, L., 1978), were randomly assigned to subjects and counterbalanced by occasion to control for topic-order effects.

Subjects were given 50 minutes to produce a completed draft under the supervision of the experimenter. Dictionaries were not available, although subjects were told to revise papers if time allowed. Instructions to subjects included the following: 1) purpose of the study, 2) time limit, and 3) writing prompts. The directions to students and writing prompts appear in Appendix A. Copies of the writing samples were returned to the participating teachers who graded the essays as part of a classroom assignment.

Dependent Measures

There were four dependent measures for each subject, i.e., mean scores over two topics and two raters from each of the four systems. The four scoring systems chosen to evaluate writing samples were selected to meet three criteria.

1. The system is a widely accepted strategy for judging writing samples.
2. The system has been used in previous studies of writing skill.
3. It represents an alternative approach to the evaluation of writing, i.e., impressionistic, analytic, or frequency counting.

General Impression Scoring (GI)

The impressionistic scoring system used in this study, developed at CSE (Smith, L.S., 1978), was adapted from procedures documented by the Educational Testing Service and the College Entrance Examination Board to score the English Composition Test (Conlan, 1976). The assumption behind impressionistic scoring is that each of the factors involved in producing a composition is related to all others, thus an overall impression is the most valid measure of writing skill. The system requires raters to read a composition through once, quickly and assign a qualitative mark. The rubric for GI appears in Appendix C1.

Procedures for rater training in impressionistic scoring follow: Raters were given materials explaining the purpose of GI scoring, the definition of exposition, copies of the topics and the scoring rubric (See Appendix C1.) Raters read the materials and discussed the rubric in terms of what exemplary essays might contain in order to conform to each of the scores on the rubric. Raters then took from their rater's packets a practice paper and scored it according to the rubric. Scores on this sample paper were reported and discussed to reach agreement on the meaning of each of the six impressionistic scores. As a result of the discussion, the rubric was amended. A copy of the amendment is included in Appendix C1. Practice papers were then scored one at a time and discussed until the raters felt that they could assign the same score as the other members of the group. At that time (after ten papers), raters read three more practice papers without discussion and compared marks. Raters then determined that they were assigning scores fairly consistently with each other and asked to have their reliability checked. A pilot test for rater reliability was conducted to assess rater consistency. Raters were given packets containing 24 sample papers consisting of both writing samples from three randomly-selected members of each of the four criterion groups. These papers did not appear in the final study. During the pilot test, reading was interrupted whenever it was noted that the ratings on sample papers diverged by more than one point. Raters were told that there was disagreement in the assignment of marks and to reread the paper. If the new set of ratings was also widely divergent, raters were told that they were grading "high" or "low". If confusion still existed as to the cause of the divergence in marking, pilot testing was stopped and the particular paper discussed. Interruptions during pilot testing added considerably to the amount of time spent in rater training and to the average time spent per paper.

Results of the pilot test indicated that further training was needed to obtain acceptably reliable scores (standardized item alpha of .80 or better). Furthermore, it was found that the papers in one group, College Low, were receiving the most divergent scores. Thus, retraining was conducted on a set of papers from College Low writers. A second pilot test for reliability was run on twenty-four (24) new sample papers, selected in the same manner as those for the first pilot test. Results of this second test indicated that rating was reliable enough (standardized item alpha=.97) to proceed to the scoring of the writing samples. Pilot test reliabilities data appear in Appendix, Tables D2-5.

The training period for impressionistic scoring lasted four hours, including two twenty minute breaks. After a break of one hour for lunch, the twenty-four (24) sample papers in the first pilot test were scored. Reading time for pilot test papers was 3-4 minutes per paper. Two hours were spent in retraining of raters and another two on the second pilot test. Reading time per paper on the final set of writing samples averaged two minutes per paper.

Diederich Expository Scale (DES)

The ETS Composition Scale (Diederich Expository Scale) represents an attempt to create an analytic scale which reflects the criteria people use when judging a piece of writing impressionistically. The scale was developed at the Educational Testing Service after a factor analytic study of the reasons judges gave for assigning grades to essays. Five factors emerged from the study and are represented on the scale as eight analytic subscales. Generally, the scores of interest from the DES are three, the subscales, General Merit and Mechanics, and the Total score (Diederich, 1974). While the entire eight category scale was used by the raters only the total scores were of interest in this study. A copy of the ETC Composition Scale (DES) appears in Appendix C2.

The training session for the Diederich Expository Scale was held one week after that for General Impression grading. Training procedures were the same as those used in the first session (see agenda appended in B) and the same four raters participated in this second session (see Figure 1, p.27 "Raters Participating in Training Conditions"). Training differed from that of GI training in that the raters spent more time clarifying the meanings of the descriptions for each of the marking categories for the DES. During the session several problems using the scale were encountered by the raters. Examples of these problems are reflected in the following rater comments: "There's no place in the rubric to penalize an off-the-topic essay." "The scale one assumes a five paragraph model essay." "Some of the descriptors in the rubric are pejorative (e.g., Under "Flavor" a "Low" is described as, 'His thought, and ideas are those of an uneducated person who does not realize how bad they sound.') "The rubric assumes you can infer the writers' motives (e.g., Under "Ideas" a "Middle" rating is characterized as, 'He writes what he thinks will sound good, not what he believes or knows.') "The rubric does not work well with short essays because often the student doesn't write enough for one to see if the descriptors fit the piece of writing". From these comments and difficulties encountered in applying the rating system to the practice papers, the raters prepared an abbreviated DES rubric which appears in Appendix C2.

Only one pilot test for reliability was necessary before raters were allowed to score the writing samples. The same twenty-four (24) sample papers used in the GI pilot test were used. Results, which appear in Appendix D3, yielded a standardized item alpha of .82.

Training for DES lasted four hours, including two 20 minute breaks. Pilot testing, due to the complexity of the system, lasted four hours and the reading time per paper averaged about 5 minutes. The rating of the writing samples proceeded more rapidly, with an average reading time of 3 minutes per paper.

CSE Analytic Scale

The CSE Analytic Scale was developed at the Center for the Study of Evaluation, University of California at Los Angeles (Smith, L. S., 1978). The scale incorporated rhetorical elements currently emphasized by teachers of composition. These elements were identified through a review of both rhetorical theory (Sanderson, 1969) and pedagogical practice (Warriner, 1973). The scale focuses on six aspects of composition: essay focus, essay development, essay organization, paragraphing, support, and mechanics. A copy of the system appears in Appendix C3.

Four raters were trained in the use of the scale, two who had learned both the GI and DES systems and two new raters. As in the previous training sessions, problems in using the scale surfaced very quickly. These difficulties included trouble in reaching consensus on the rating categories for the "Mechanics" section and confusion as to the meanings of the categories, "Focus", "Development", and "Organization". Raters were also uncomfortable with the fact that the scale had no provision for judging such aspects of composition as style and originality.

The pilot test for CSE reliability, results appended in D4, was run on the set of twenty-four (24) sample papers described in previous sections. A standardized item alpha coefficient of .96 was obtained and raters proceeded to the scoring of the final papers.

Training time for CSE was approximately six hours, distributed over two evening sessions. Pilot testing consumed approximately two half hours with an average of 5 minutes spent per paper. Rating time dropped to 3 minutes per paper on the final reliability test.

T-Unit Analysis

The term T-unit refers to a minimally terminal syntactic unit, i.e., the shortest gramatically complete sentences that a passage can be cut into without creating fragments (Hunt, 1965). T-units are single main clauses and all subordinated clauses that can be terminated grammatically with a period or other terminal punctuation mark. In contrast to impressionistic scoring, which attempts to model human judgment, or analytic systems, which try to highlight the salient characteristics of "the essay", T-unit analysis focuses on the syntactic features of a peice of writing. The concern is not with the unity of a writing sample, not with the presence of rhetorical elements, but with the level of syntactic "maturity" of a piece. Syntactic maturity, or the mean length of T-units, is an indirect measure of the language development of the writer. This method of analysis is often dubbed "quantitative" (Howerton, 1977; O'Hara, 1973) to distinguish it from judgmental models of composition evaluation.

The procedure of performing T-unit analysis is as follows: Raters segment the writing sample into T-units (main clauses and all embedded clauses). Raters then count the number of words in the composition and divide them by the number of T-units. The result is the T-unit score, the mean number of words per T-unit. The higher the mean number of words per T-unit, the more "systematically mature" is the piece of writing. From previous studies (Hunt, 1964, 1965, 1970) it was expected that mean scores in this study would range from a low of seven to a high of 20 words per T-unit.

Raters who were trained on the DES also trained in T-unit analysis. Raters were given the segmenting rules (see Appendix C4) and practice papers and were told to use the rules to find the mean number of words per T-unit. No changes were made in the rules by the raters.

The major difficulties encountered by the raters in using T-unit analysis were boredom and the inability to accurately count the number of words in the essays. Pilot testing for reliability was conducted on the 24 sample papers used in the previous pilot studies. A standardized item alpha coefficient of .98 was obtained for T-unit.

T-unit analysis consumed the least amount of training time, about one hour, but the most amount of scoring time. Time spent on each paper was about 2 minutes but rater fatigue and boredom necessitated frequent breaks. Rater comments during scoring indicated that the system was not only tedious but "purposeless".

ANALYSES

A prerequisite to the analysis of writing performance was the examination of the reliabilities and validities of the scoring systems used in this study. Interpretation of results bearing on the relationship of scoring criteria to writing skill was dependent on comparative reliabilities of the scoring systems and whether the systems measured the same or different aspects of writing performance. The choice of which scoring systems are most appropriate for measuring the writing skills of high school and college students assumes that the systems selected are acceptably reliable and that the constructs they measure are known.

Rater Reliability

It was hoped that the controls on writer, topic, and test occasion variation, described in the previous section, in conjunction with rater training on each of the scoring systems, would lead to comparable reliabilities for the different systems. The actual reliability coefficients did meet the criterion for rater reliability, i.e., alpha = .80 or above.

There was, however, some variation in reliability among systems with the analytic and frequency counting methods more reliable after training than general impression. The alpha coefficients were: General Impression .81, Diederich Expository Scale .90. CSE Analytic Scale .97 and T-unit analysis .99. These coefficients represent an average over two writing samples per examinee and two ratings per sample. Table 2 displays the analyses of variance from which the intraclass correlations were computed. The formula used was $\frac{MS_p - MS_r}{MS_p}$. Two sources of random variation, subject and rater were specified; topic was considered a fixed factor in the analysis. Although the complete design for accounting for variation associated with writing samples considers topic, occasion, rater and subject as random factors, pilot test results for rater reliability (see Table 1 Appendix D) indicated that an interaction existed between topic and subject. The finding that some students write well on one topic but not the other suggested that the topics were not random samples from the domain of exposition, or that the definition of the "domain" was too broad. When topics are not of equivalent difficulty, the examinee score of interest is the average over the two topics, a practice that does not allow one to generalize across topics for reliability estimates.

Scale Intercorrelations

The second question bearing on the interpretability of results is: Are the scoring systems measuring the same constructs? The assumption that the selected essay scoring criteria represented different conceptualizations of the writing process was tested by examining the inter-correlations among systems. The data were analyzed twice, once using the entire sample (N=80) and a second time by criterion group (High School Low, High School High, College Low, and College High).

The results are contradictory. Intercorrelations among three of the scoring systems, General Impression, the Diederich Expository Scale and the CSE Analytic Scale are quite high for the total sample (Refer to Table 3). It appears that across the entire sample these systems measure elements in common. T-unit analysis alone stands apart as an alternative approach to measuring writing skill.

However, the relationships among the three highly correlated systems change dramatically when the data are analyzed by criterion group. A review of Table 4-7 indicates that for two groups, High School Low and College High, the relationships among General Impression, the Diederich Expository Scale, and the CSE Analytic Scale are still relatively high. Intercorrelations among the three "related" systems are weakened for the High School High and College Low writers. These results suggest that for some writers, notably High School Low and College High, the distinguising aspects of writing performance may be those shared by three of the scoring systems, while for other writing groups, writing performance is characterized by elements belonging to only one or two of the scoring systems. The possibility of range restriction on performance within certain criterion groups cannot be dismissed. However, an examination of the means and standard deviations on Table 8, as well as the distributions of scores displayed in Figures 7-10, shows that range restriction is not a major problem.

T-unit analysis is uncorrelated with the other systems when both the entire sample and the criterion groups are the unit of analysis. In fact, T-unit scores are negatively correlated with the other systems for the two college writing groups. The lack of relationship of T-Unit to either General Impression marking or the Diederich Expository scale was unexpected in light of the findings of the O'Hare (1973) and Howerton (1977) studies.

Results of the correlation analysis suggest that for some kinds of writers, writing performance may vary according to the set of scoring criteria employed. They also indicate that the four systems do represent different conceptualizations of the writing process when they are used to assess the writing performance of relatively homogeneous groups of students.

The Performance of Criterion Groups Across Scoring Systems

Two questions central to this study relate to the writing performance of criterion group members as it is evaluated by four different scoring systems: 1) What is the pattern of performance of each criterion group across the four scoring systems? and 2) What is the pattern of performance of different criterion groups on the same system? Descriptive statistics for group performance on each of the systems are displayed in Table 8. As the criterion groups were chosen to represent four levels of writing skill, it was expected that the mean performance of each group would rank as follows: High School Low, High School High, College Low, and College High. There were two unexpected findings: 1) The College Low group outperformed the College High group on three of the four scoring systems (General Impression, the Diederich Expository Scale and the CSE Analytic Scale) and 2) T-unit results did not rank the groups as expected. Both the high school and college "high" groups did write longer T-units than the low groups at each level, but there is no trend toward longer T-units as assumed writing skill increases.

A comparison of criterion group standard deviations suggests that as assumed writing skill increases, variability in group performance increases. T-unit again is the exception. Score dispersion around the mean was least for General Impression and greatest for the Diederich Expository Scale. However, an examination of the actual score distributions (Figures 3-6) suggests that this wide dispersion may be more a function of the DES score scale, ranging from 5-45 points, than any real difference in the tendency of raters to assign scores differently with different scoring systems.

The performances of different criterion groups on the same scoring are displayed in Figures 3-6. For General Impression marking, three of the groups have similarly shaped score distributions: The High School Low, High School High and College Low groups. When the Diederich Expository Scale is used to assign essay scores, only the High School and College Low groups show similar shapes. Score distributions of the criterion groups are dissimilar for the CSE Analytic Scale; when T-unit analysis is used, the High School High, College Low and College High Groups have similar shapes and locations.

A comparison of the score distributions of the criterion groups on the same scoring system shows no consistent pattern of score assignment across groups for any one system. Figures 7-10 display the pattern of performance of the criterion groups on the four systems. The High School Low group's scores are assigned in the same pattern for three of the scoring systems, General Impression, the Diederich Expository Scale and the CSE Analytic Scale. The score assignments do appear to be most consistent for the lowest writing group and least consistent for the highest writing group. Distributions are most skewed for General Impression for College High, the Diederich Expository Scale for High School High, and CSE for College Low. Clearly with the exception of High School Low, the distributional shapes change across scoring systems.

The Efficiency of Scoring Systems for Predicting Criterion Group Status

Research questions three and four explore the efficiencies of the scoring systems for predicting criterion group membership: The questions of interest are: 3) Which scoring systems discriminate most clearly among criterion groups? and 4) Which systems are the best predictors of criterion group status? The investigation of these issues was predicated on the assumption that the scoring systems, when considered separately, were able to distinguish among criterion groups. This assumption was tested with a one way analysis of variance. The results of the one way ANOVAS are displayed in Tables 9-12.

Group differences were evident on three of the systems, General Impression, the Diederich Expository Scale, and the CSE Analytic Scale. The Eta squared statistics accompanying the tables suggest that for these three systems, a large proportion of the variation among examinee scores is associated with differences among criterion groups. The one way analysis of variance results also suggest that significant effects would be found for scoring systems in subsequent multivariate analysis and that T-unit analysis would have little contribution to the discrimination among or accurate classification of examinees.

Multivariate Analysis of Variance

A multivariate analysis of variance was performed using group membership as the independent variable and scores on all four scoring systems as dependent variables in order to determine which systems discriminated best among criterion groups. The multivariate analysis of variance null hypothesis of equality of main vectors was rejected with an F of 1530.599 (df 4,473 $p < .0001$). Post hoc comparisons (refer to Table 13) revealed significant differences between the means of the High School Low and each of the other groups and between the High School High and College Low groups over all systems. No significant differences were found between the College High and either the High School High or College Low groups.

Two discriminant analyses were run to investigate the relative contribution of each of the scoring systems to group separation. The scores for T-unit analysis were dropped from the discriminant analyses because preliminary results indicated that there were no significant differences among the mean T-unit scores of the criterion groups. The first analysis investigated the effects of using scores for General Impression, the Diederich Expository Scale and the CSE Analytic Scale to separate groups. The second examined the effects from the DES and CSE scales alone. For each analysis, one significant Linear Discriminant Function (LDF) was obtained (Refer to Tables 15-16). Two kinds of data were examined to identify those scoring systems contributing most to group separation, the within groups correlation matrix and the standardized coefficients of the LDF's. The within groups correlation matrix, displayed in Table 14, indicates that the Diederich scale combined with either the CSE scale or General Impression should contribute most to the discriminant functions separating the criterion groups. An examination of the standardized coefficients (Tables 15-16) of the significant LDF's supports the hypothesis that DES and CSE are the "strongest" variables.

Discriminant Classification

The ability of the systems to classify writers into the correct criterion groups was also investigated with discriminant analysis. For this analysis, the scoring systems became the independent variables and criterion groups the dependent variables. Results were obtained in the form of classification tables reporting the percentage of examinees falling into each of the four criterion groups if first three systems, GI, DES, and CSE, and then two systems, DES and CSE, were used. Tables 17-18 present the percentage of cases that the classification function would assign to each writing group. The results indicate that the most accurate classification functions are those derived from GI, DES, and CSE. This finding is consistent with the finding that the greatest separation/discrimination among criterion groups was obtained using the scores on three, rather than two of the systems.

The misclassification errors themselves have an interesting pattern. Few High School Low writers are assigned to the next highest writing group and none are assigned to college level writing groups. No High School High writers are assigned to the High School Low group, but more are assigned to the College High groups than to the next highest group, College Low. Misclassification errors for the College Low group are in the direction of lower rather than higher ability writers. And, the poorest classification accuracy is for the College High writers. An alarming 10% of this group are misclassified into the High School Low group, a group whose writing scores are clearly discriminable from the others and whose mean performance is significantly poorer for the three scoring systems used in the classification analysis.

If two scoring systems, DES and CSE, are used in the classification functions, classification accuracy drops from 61.25% to 58.75%. Misclassifications occur with greater frequency for the two college groups when only the two analytic systems are used to predict group membership. The direction of misclassification is of note: The loss of classification accuracy when only two scoring systems are used accrues to the College Low and College High groups alone. That is, there is an increase in the number of College Low students misclassified as College High and visa versa. It is possible that the two analytic systems are not "picking up" some essential differences between the two writing groups that is accounted for by General Impression scoring.

Summary of Results

Several preliminary analyses were necessary in order to select the appropriate procedures for analyzing data related to the four questions central to the present study. Rater reliabilities were calculated. The relationship among scoring systems was examined with Pearson correlations. Finally, univariate analyses of variance for each scoring system were run to identify those systems able to distinguish among writing groups. The results of these analyses indicated that, while T-unit was the most reliable system, it was least associated with aspects of writing performance measured by the other systems. When T-unit analysis was used, group differences did not account for a large amount (only 2%) of the variation in examinee performance. Rater reliabilities also indicate that the two analytic systems, DES and CSE were capable of producing more homogeneous ratings after training than the general impression system chosen. Intercorrelations among the three systems having the greatest effect on writing performance, GI, DES, and CSE, at first appeared high. However, when the data were analyzed by criterion group, the relationship among these systems fluctuated from mild to weak. The appearance of significant group differences for the general impression and analytic systems in the univariate analysis indicated that the response criteria were able to detect group differences and that data analyses related to the four research questions could be run.

Two research questions focused on the patterns of performance of criterion group members on scoring systems. Those patterns are displayed in Table 8 and Figures 3-10. An unexpected finding related to group performance was that, with the exception of T-unit analysis, mean scores of the College High group were lower on all systems than mean scores of the College Low group. No consistent patterns of performance were found for any one group across systems or for any one system by group. However, the High School Low writers did have the most similarly shaped distributions of scores across systems, and T-unit analysis the most similarly shaped score distributions across groups. The descriptive statistics and graphs indicate that, overall, the scoring systems are measuring different aspects of writing performance of the criterion groups with the implication that, for some groups, one system may be better able to characterize writing performance than another. There are, then, group by scoring systems interactions worthy of investigation.

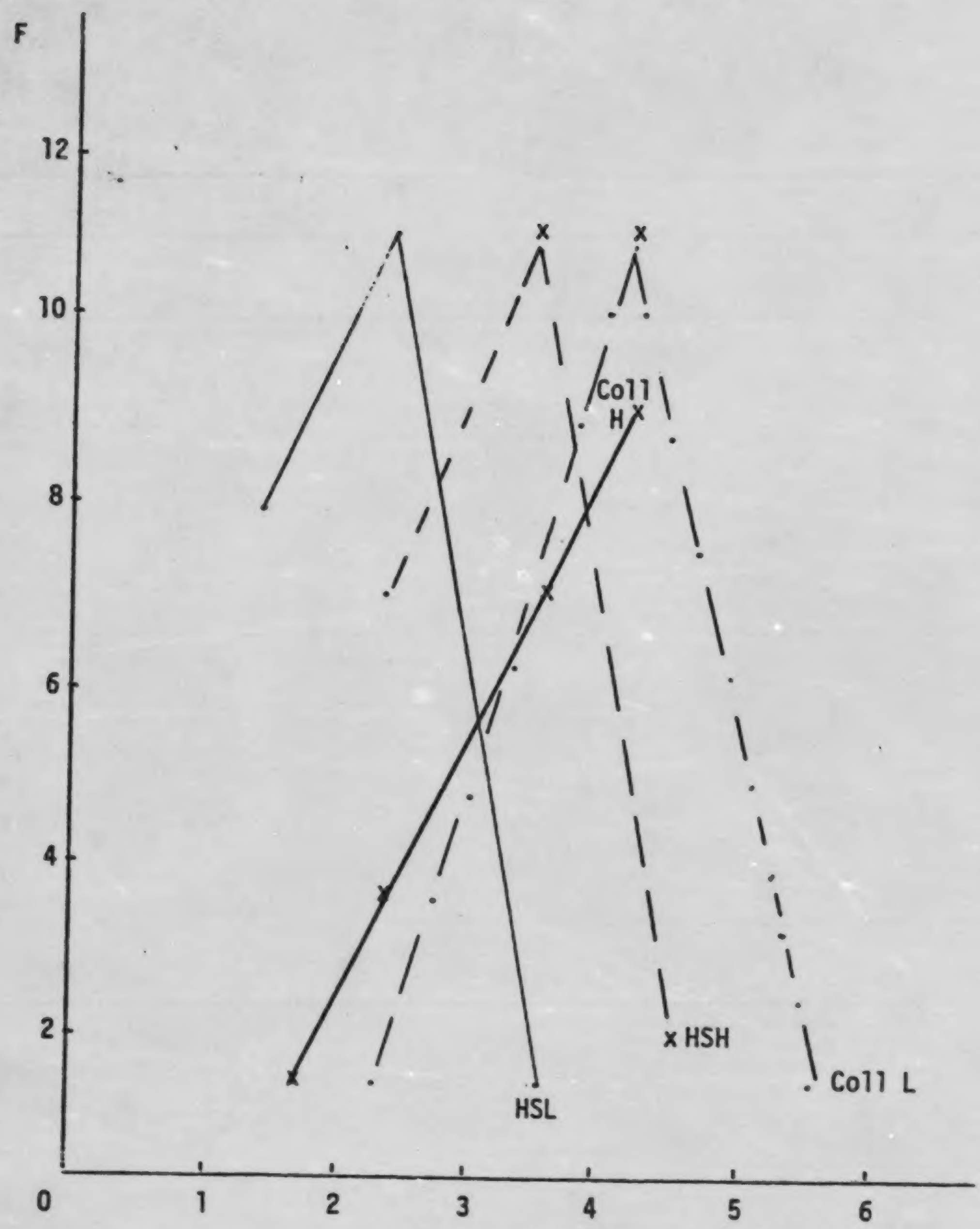
A final finding of the preliminary analyses was that T-unit analysis did not measure what is conventionally described by either holistic judgments or currently popular analytic criteria "good expository skill". Its negative relationship to the other systems for some of the criterion groups suggest that for the present study, the information gained from T-unit scores would be of little benefit in describing or predicting expository writing performance for writers of different abilities.

. Two research questions were concerned with the problem of identifying those scoring systems that best discriminated among criterion group members and classified them into the correct groups. Results from the preliminary analyses indicated that multivariate analyses would be most appropriate to investigate these questions. A multivariate analysis of variance (MANOVA) was run to determine if group differences did indeed exist when all four scoring systems were considered simultaneously as independent variables. The significant MANOVA indicated that the null hypothesis of equivalent mean vectors was not supported and thus a discriminant analysis could be performed. All groups were separated significantly with the exception of the College High from the High School High and College Low groups.

Two discriminant analyses, each using different combinations of scoring systems, were run to determine which systems were contributing the most to group separation while at the same time capable of producting discriminant functions with a high classification accuracy. One significant discriminant function was found for each combination of scoring systems: GI/DES/CSE and DES/CSE. An examination of the within group correlation matrices and standardized Linear Discriminant Function coefficients, indicated that the DES contributed most to group separation when combined with either CSE or GI. Both DES and CSE contributed more to discrimination than GI.

The ability of the scoring systems to accurately predict group membership was determined by examining the classification tables produced in the two discriminant analyses. It was found that the use of three scoring systems

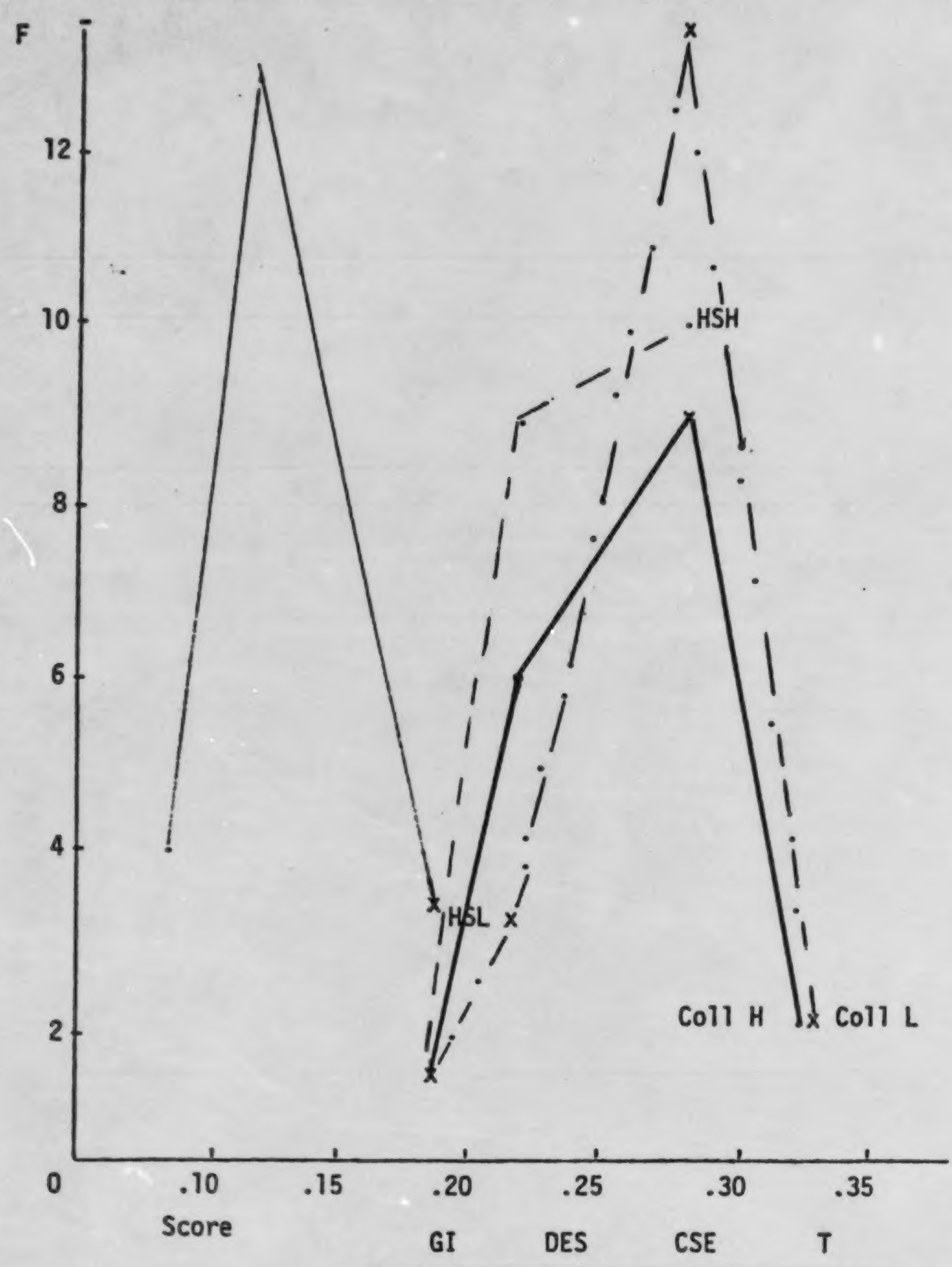
(GI, DES, CSE) produced more accurate classification results than using scores on only DES and CSE. It was also found that classification accuracy was greatest for the High School Low group and least for the College High group. Prediction was much better for the poorer writers than the better ones. The loss of classification accuracy when only two analytic systems were used was minimal and affected only the college groups. The pattern of predictions, accurate for High School Low and College Low, fair for High School High, and abysmal for College High, remained the same despite the combination of scoring systems used.



	HSL	HSH	Coll L	Coll H
$\overline{X}$	1.97	3.21	4.01	3.33
S.D	.50	.65	.72	.74

FIGURE 3

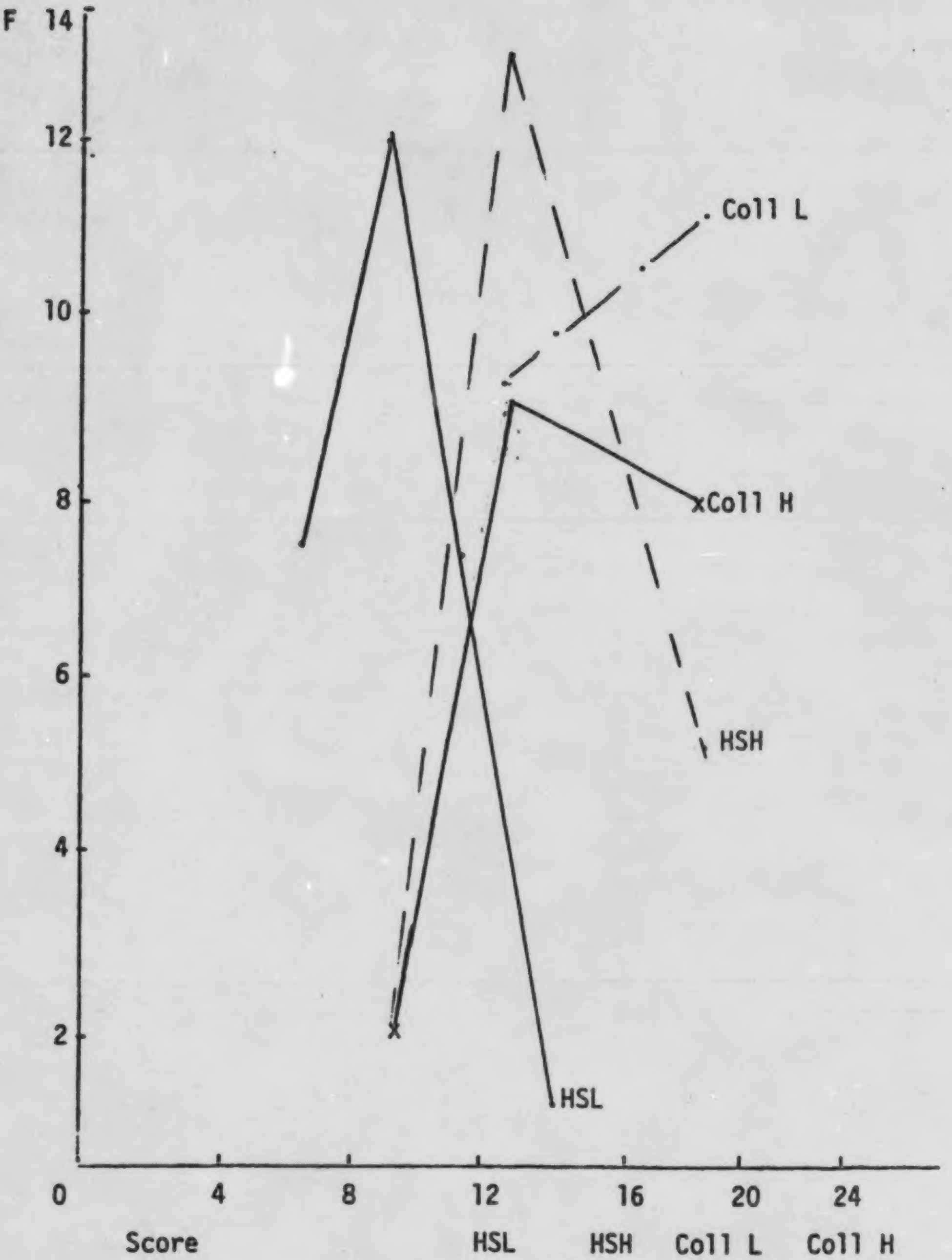
General Impression: Comparison of Score Distributions Across Criterion Groups



	GI	DES	CSE	T
$\bar{X}$	13.25	25.66	28.00	26.00
S.D	2.95	2.91	3.26	409

FIGURE 4

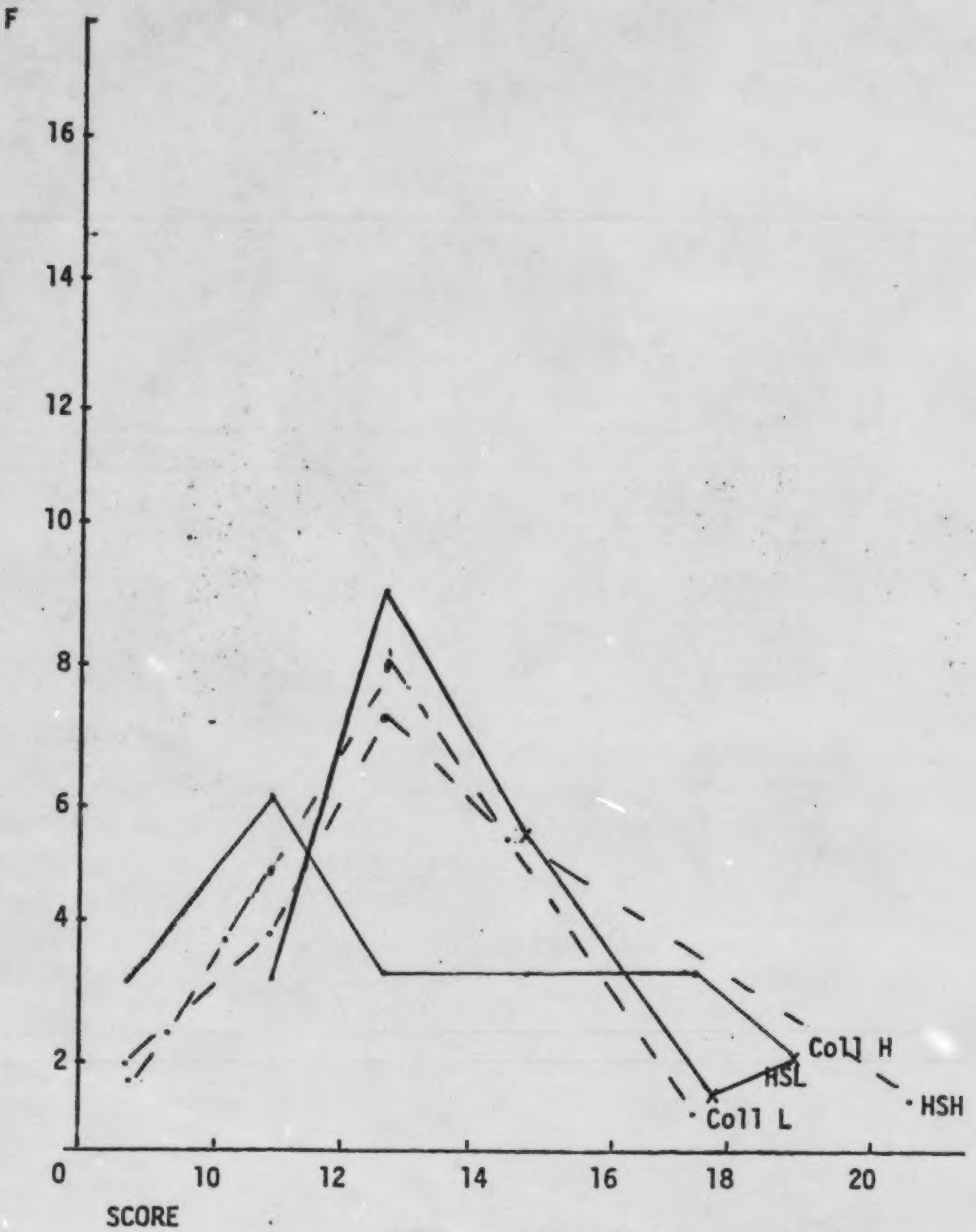
Diederich Expository Scale: Comparison of Score Distributions Across Criterion Groups



	HSL	HSH	Coll L	Coll H
$\overline{X}$	9.84	15.50	17.07	16.18
S.D.	1.57	1.92	2.08	2.83

FIGURE 5

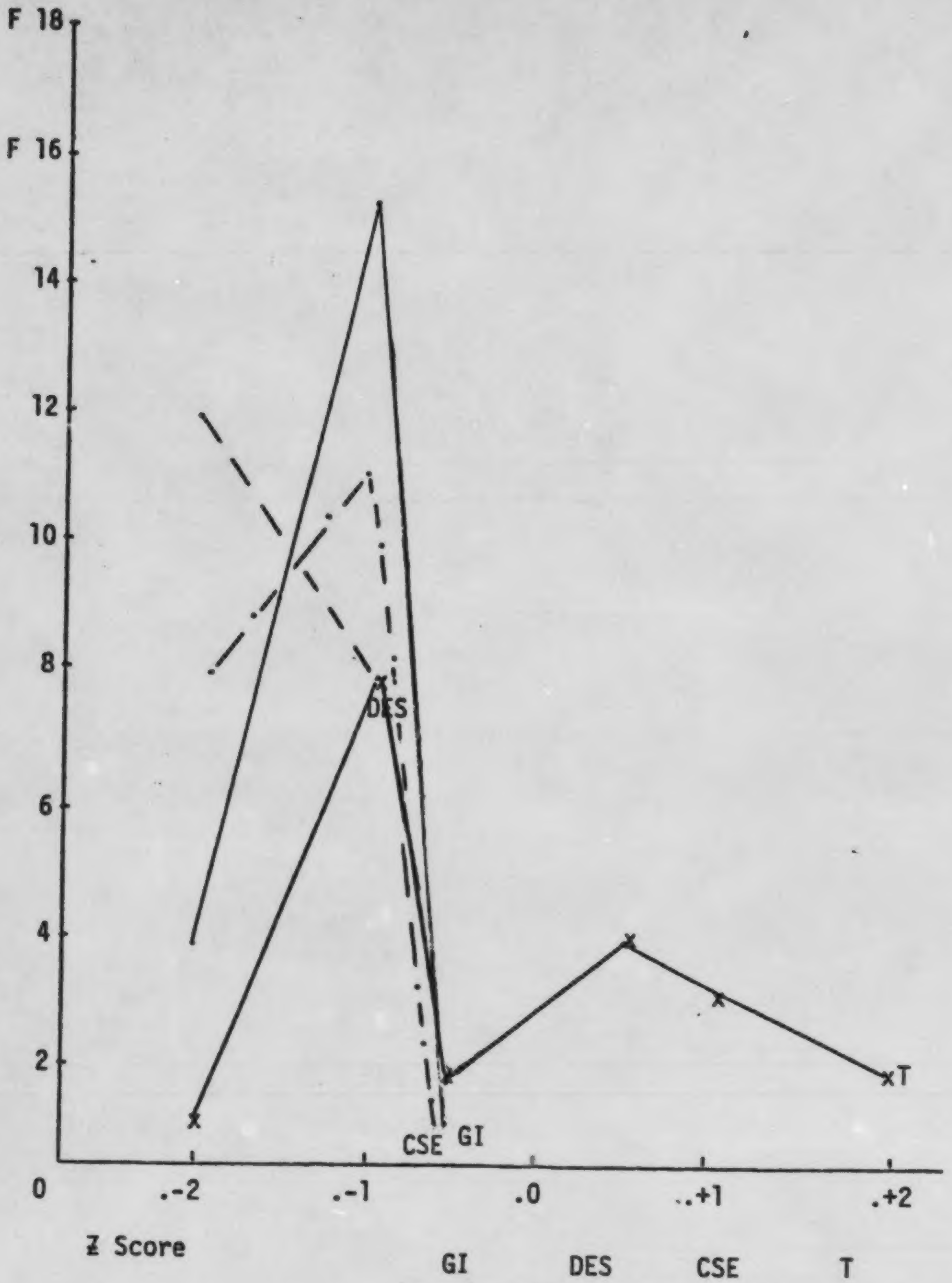
CSE Analytic: Comparison of Score Distributions Across Criterion Groups



	HSL	HSH	Coll L	Coll H
X̄	14.18	15.07	13.99	14.89
S.D.	3.45	5.18	1.95	2.06

FIGURE 6

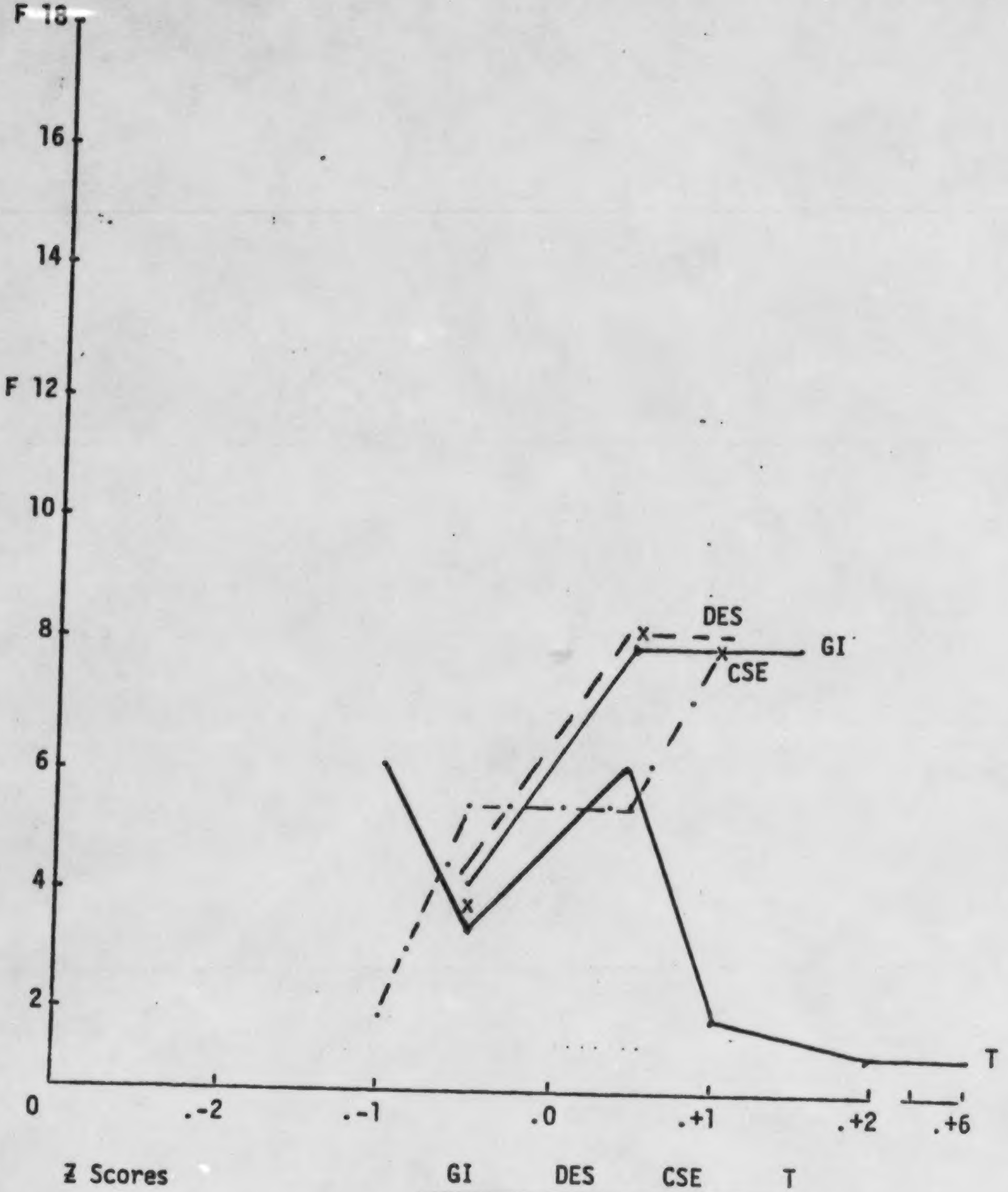
T-Unit: Comparison of Score Distributions Across Criterion Groups



	GI	DES	CSE	T
X̄	-1.2	-1.4	-1.3	-0.1
S.D	.50	.44	.44	1.0

FIGURE 7

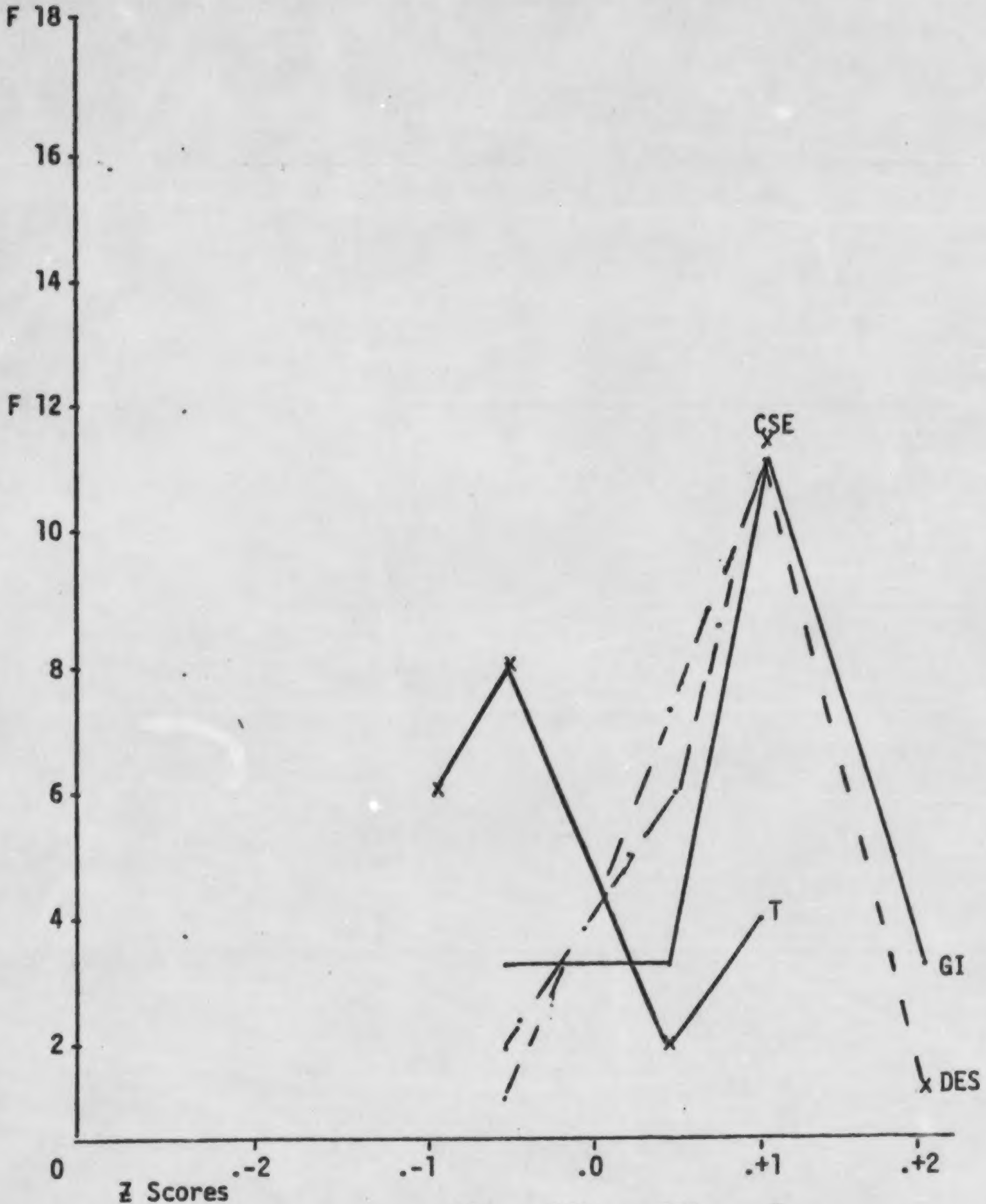
High School Low: Comparison of Z Score Distributions Across Scoring Systems



	GI	DES	CSE	T
X	.35	.36	.24	.16
S.D	.43	.44	.54	1.5

FIGURE 8

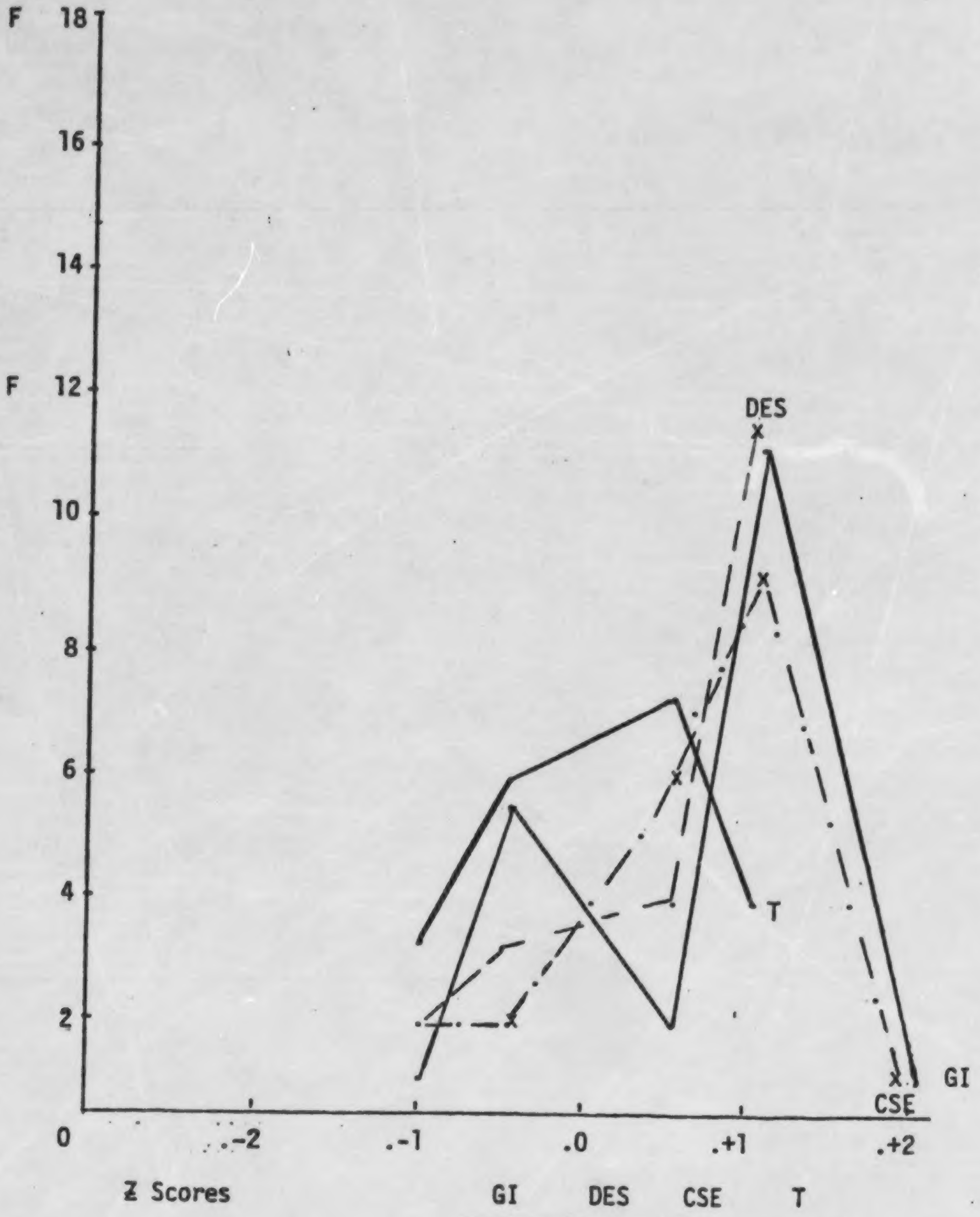
High School High: Comparison of Z Score Distributions Across Scoring Systems



	GI	DES	CSE	T
$\overline{X}$	.81	.70	.68	-.16
S.D	.72	.49	1.4	.38

FIGURE 9

College Low: Comparison of Z Score Distributions Across Scoring Systems



	GI	DES	CSE	T
$\overline{X}$	.44	.42	.43	.11
S.D	.74	.61	.80	.61

FIGURE 10

College High: Comparison of Z Score Distributions Across Scoring Systems

TABLE 2

Data for Estimating Rater Reliabilities of Scoring Systems
(Average over two writing samples per student
with two raters per sample)

Scale Name	Number of Students	SS (Students)	SS(Raters)	SS(Students x Raters)
General Impression	160	390.888	0.0	74.000
Diederich Expository Scale	158	17336.218	3.041	1667.459
CSE Analytic	156	12918.419	0.028	330.471
T-Unit	160	3677.801	0.053	29.619

TABLE 3

Intercorrelations Among Scoring Systems

	Total Sample n=80			
System	GI	DES	CSE	T-Unit
GI	-			
DES	.82	-		
CSE	.79	.86	-	
T-Unit	.004	.06	.001	-

TABLE 4

Intercorrelations Among Scoring Systems

	High School Low n=20			
System	GI	DES	CSE	T-Unit
GI	-			
DES	.64	-		
CSE	.71	.79	-	
T-Unit	.28	.33	.30	-

TABLE 5

Intercorrelations Among Scoring Systems

High School High
n=20

System	GI	DES	CSE	T-Unit
GI	-			
DES	.54	-		
CSE	.44	.30	-	
T-Unit	.003	.24	.21	-

TABLE 6

Intercorrelations Among Scoring Systems

College Low
n=20

System	GI	DES	CSE	T-Unit
GI	-			
DES	.53	-		
CSE	.11	.37	-	
T-Unit	-.07	-.39	-.52	-

TABLE 7

Intercorrelations Among Scoring Systems

College High
n=20

System	GI	DES	CSE	T-Unit
GI	-			
DES	.51	-		
CSE	.70	.69	-	
T-Unit	-.31	-.23	-.52	-

TABLE 8

Criterion Group Performance Across Scoring Systems: Means and Standard Deviations

SYSTEM	Group HS Low n=20	HS High n=20	Coll Low n=20	Coll High n=20	Total n=80
GI (Total 6)	$\bar{X}$ = 1.97 S.D.= .50	3.21 .65	4.01 .72	3.33 .74	3.02 1.00
DES (Total 45)	$\bar{X}$ =13.25 S.D.= 2.95	25.66 2.91	28.00 3.26	26.00 4.09	23.29 6.68
CSE (Total 24)	$\bar{X}$ = 9.84 S.D.= 1.57	15.50 1.92	17.07 2.08	16.18 2.83	14.65 3.55
T-Unit (Range 10-34)	$\bar{X}$ =14.18 S.D.= 3.45	15.07 5.18	13.99 1.95	14.89 2.06	14.52 3.38

TABLE 9

Analysis of Variance: System 1
General Impression

Source	SS	DF	MS	F
Between	752.637	3	250.879	36.094*
Within	528.250	76	6.951	

ETA^2=.587

TABLE 10

Analysis of Variance: System 2 DES

Source	SS	DF	MS	F
Between	42930.338	3	14310.113	80.522*
Within	13506.550	76	177.718	

ETA^2=.760

TABLE 11

Analysis of Variance: System 3 CSE

Source	SS	DF	MS	F
Between	10280.70	3	3426.09	46.32*
Within	5622.50	76	73.98	

ETA^2=.646

TABLE 12

Analysis of Variance: System 4 T-Unit

Source	SS	DF	MS	F
Between	263.740	3	87.91	.47(NS)
Within	14205.086	76	186.90	

ETA^2=.0182

TABLE 13

F Ratio for Multivariate Tests of the Equality of Mean Vectors

Contrasts		F	DF	P
High School Low versus				
	High School High	4.750	4.73	.0019*
	College Low	21.629	4.73	.0001*
	College High	37.829	4.73	.0001*
High School High versus				
	College Low	2.934	4.73	.026*
	College High	1.183	4.73	.325
College Low versus				
	College High	1.181	4.73	.326

*Significant F ratio

TABLE 14

Results of Discriminant Analysis Within Groups Correlation Matrix

System	GI	DES	CSE	T-Unit
GI	-			
DES	.54	-		
CSE	.48	.55	-	
T-Unit	.01	.04	-.04	-

TABLE 15

Results of Discriminant Analysis,
Retained Functions Systems 1-3

Function	Eigen Value	Relative Percentage of Information Accounted for	Wilke's Lambda	Standardized LDF Coefficient
1	3.388	96.69	.204	GI - 0.061 DES - 0.758* CSE - 0.214
2	0.115	3.27	.896	GI - 1.741 DES - 1.493 CSE - 0.002
3	0.001	.03	.999	GE - 0.548 DES - 1.408 CSE - 2.039

TABLE 16

Results of Discriminant Analysis,
Retained Functions Systems 2-3

Function	Eigen Value	Relative Percentage of Information Accounted for	Wilke's Lambda	Standardized LDF Functions
1	3.374	99.69	.226	SYS 2-DES .793* SYS 3-CSE .232
2	0.010	.31	.990	SYS 2-DES -1.791 SYS 3-CSE 1.945*

TABLE 17

Classification Accuracy of the
GI, DES, and CSE Scoring Systems

Actual Group	N	Predicted Group Membership			
		HSL	HSH	Coll L	Coll H
High School Low	20	95%	5%	0%	0%
High School High	20	0%	55%	10%	35%
College Low	20	0%	25%	70%	5%
College High	20	10%	20%	45%	25%

Percent of cases correctly classified: 61.25%.

TABLE 18

Classification Accuracy of the DES and CSE Scoring Systems

Actual Group	N	Predicted Group Membership			
		HSL	HSH	Coll L	Coll H
High School Low	20	95%	5%	0%	0%
High School High	20	0%	55%	20%	25%
College Low	20	0%	25%	65%	10%
College High	20	10%	20%	50%	20%

Percent of cases correctly classified: 58.75%.

CONCLUSIONS

The purpose of this study was to identify those essay scoring systems that would have the most utility for making decisions about the writing skill of high school and college students. The scoring criteria were selected to represent three different approaches to essay grading: general impression, analytic, and frequency counting. The merit of each of the systems was evaluated according to its reliability, its ability to discriminate among examinees with different levels of writing skill and its ability to classify examinees with known kinds of writing skill into the correct criterion groups. In addition, the performance of the criterion groups on each of the scoring systems used in the study was compared to see if the use of different essay scoring criteria resulted in different patterns of examinee performance. The discussion of the findings bearing on these issues is presented in sections on the relative reliabilities of different kinds of scoring systems, the effect of differing response criteria on patterns of writing performance, and the validities of the systems for making placement decisions involving high school and college writers.

Reliability

It was expected that the reliabilities of the scoring systems used in this study would be comparably high due to the use of rater training sessions. The reliabilities were higher than expected from results of the pilot tests for reliability (See Appendix D). During training, reliabilities for the four systems were GI, .97, DES, .82, CSE, .96, and T-unit .98. Although final reliabilities showed a drop in reliability for GI from .97 to .81, there was an increase for the other three (DES to .90, CSE to .97, and T-unit to .99). It was expected that the reliabilities for the final set of writing samples would be somewhat lower as only two ratings would be used. However, the additional "practice" gained from reading more papers probably offset the expected drop in rater agreement.

The differential reliabilities among scoring systems can also be explained by two other factors: 1) training effects and, 2) specificity of scoring rubrics. The order of rater agreement (from a low to high in the following order, GI, DES, CSE, and T-unit) corresponded with the order of the training sessions. Two of the raters participated in all four sessions. The additional practice gained in "reaching consensus" as to the meanings of the various scoring rubrics may have contributed to the progressive increase in rater agreement.

The second factor bearing on the ordering of the reliabilities lies within the nature of the scoring rubrics themselves. GI is the least structured rubric followed by DES, then CSE. Unit rules are so unambiguous that persons well versed in grammar can quickly agree as to what constitutes a minimally terminable unit of thought.

That GI should have the lowest reliability is no surprise. Its rubric gives little direction to raters. For example, the "High" category is defined as, "An excellent example of exposition." In order to agree on the meaning of this statement, raters had to discuss many sample papers. This ambiguity of the rubric is further underscored by the fact that two pilot tests for reliability had to be run before an acceptable level of rater agreement was reached.

Like GI, the DES rubric uses descriptors open to various interpretations (See Appendix C2). Much training time was spent clarifying the meanings of such descriptors as, "He discusses each main point long enough to show clearly what he means." By contrast, the CSE analytic scale descriptors of "High" on the analogous category "Development" was, "All ideas and details in this paper are clearly and logically related to a single dominant 'main idea'. Development of the idea is not interrupted by digressions of thought or extraneous material." This second rubric gives much clearer descriptions for assigning essay scores.

The Effect of Differing Response Criteria on Patterns of Writing Performance

The writing performance of each of the criterion groups was examined by looking at the shapes of the distributions produced by the four scoring systems. Score distributions varied more, however, for the high writing groups than the low. No scoring system evidenced the same pattern of performance for all four groups, although both General Impression and T-unit produced similarly shaped distributions for three of the groups. The lack of similarity in distributional shapes for the groups on the two analytic systems is a surprise, as these systems were thought to represent a similar conceptualization of the writing process.

An unexpected finding about criterion group performance across systems is that, although the groups were selected to represent four different levels of writing skill (high school low performance and high performance, as well as college low and high performance) a comparison of mean scores of groups on the four systems does not support this assumption. For three of the scoring systems, GI, DES, CSE, the College Low group "outperforms" the College High--a group whose writing performance is "better" by virtue of having passed a written test of English composition.

That the findings do not indicate a clear superiority for the College High group may be due to several factors: First, the examinees in the College Low and High groups may be a non-representative sample. The classes selected to participate in this study met during a summer session. There is some evidence that students taking summer session classes have different writing backgrounds from those in regular session; the delay in taking the composition course is often attributable to carrying large math or science course loads during the regular session.

Second, many examinees in the College Low group were foreign-born. Although they failed the initial composition screening examination, their writing skills may not have been "poor" on the structural dimensions measured by the scoring systems of interest in this study. That this hypothesis has some credibility is born out by the fact that only two of the scales can detect awkward language; GI as part of the overall impression grade and DES in the category "Flavor". Thus, it would be possible for examinees who wrote mechanically correct, albeit awkward sentences, to score adequately on the CSE analytic scale or T-unit. Additionally, awkwardness, the use of non-English constructions, and lack of felicity in phrasing are not explicitly measured by GI or DES, the first of which is concerned with the overall message and the second of which amply regards correct mechanics and presentation of ideas.

A third explanation of the reversal in means of the College Low and College High groups is that these groups were not terribly different in writing composition skills because the initial screening examination was unable to detect real differences. Students are placed in these classes on the basis of only one writing sample. Also, those aspects of writing performance that were assessed by the scoring systems in this study may not have been thought crucial to the Subject A examiners. The relatively small number of subjects in the study presents another alternative explanation. Finally, the possibility of misplacement of examinees by the Subject A examination exists. It is unclear whether the Subject A examination is valid for the purpose it is currently being used for.

Validity

The underlying purpose of this study is to identify those scoring systems best able to differentiate among different levels of writing competence. The central issue investigated was: Do decisions about writing competence depend on the scoring procedure selected? It can be concluded that for the four systems of interest in this study, the answer is, "yes". Different systems do produce different patterns of results, which may mean that the writing performance of each group cannot be adequately described by the same scoring system. The differential results of the distributions of scores on the four systems may indicate that the writing skills of the criterion groups of interest in this study are essentially different.

The variation in performance of criterion groups across scoring systems found may be explainable in terms of the elements featured by each scoring system. High scores on General Impression, for example, indicate that the writing is characterized by good organization of thought, unity, and has a distinctive style. High Diederich Expository scale ratings are earned by providing a clear organization, substantial support and having a command of English mechanics. High CSE essays tend to follow the classic five paragraph essay model. T-unit scores, on the other hand, are associated only with the length and complexity of the sentences employed. It was expected that General Impression scores would be appropriate for all groups. The results partially substantiate this expectation; it is perhaps only the unusual performance of the College High group which causes this system to perform differently for the four groups.

The DES may be better for discriminating between the writing samples of Low and High groups because of its emphasis on mechanics and organization. These problems are frequently encountered in poor writing. The CSE scale appears to produce more regular patterns of performance for high school than for college groups. It is possible that the heavy reliance of the five paragraph model essay, taught in high school and mastered by college, in this scale is responsible for the differences in patterns of performance.

By college, other aspects of writing other than organization and support may separate the effective from the ineffective writers.

The fact that T-unit was found to be useless as a discriminator among levels of writing skill was expected. T-unit is a measure of syntactic maturity, the ability to use complex language patterns. Although previous studies with syntactic maturity have found that increasing T-unit length is associated with higher "quality" writing, these results were obtained with groups representing a wide age range (Hunt, 1970). The relative maturity and homogeneity of the present sample may account for the lack of discriminatory power for the T-unit. The expected average T-unit length for high school seniors is 14.4 (O'Hare, 1973). The actual T-unit average of the groups in this study ranged from 13.9 to 15.07, very close to the mean results of the earlier studies of syntactic maturity. The lack of significant differences among the groups on mean T-unit length may be more a result of the fact that syntactic maturity has been reached by the writers in this study than that T-unit is unable to detect differences among writers over a wide span of ages.

Classification

Test results are often used to make decisions about examinee competence either to graduate from high school or to continue on to the next level of instruction. These decisions have instructional impact in that once students are classified as competent or incompetent, appropriate remedial instructional programs must be provided. At the high school level the competence decision assumes that students who pass the examination have "mastered the subject" and are indeed able to perform like-tasks after graduation. In college, this decision is more often associated with instructional placement: Students failing a writing test are placed in remedial composition programs while those who pass are assigned to regular programs.

In this study, competent examinees were defined _a priori_ as members of the "high" criterion groups. Competent high school students were identified by teacher judgment, as reflected in English composition grades and admission to the pre-college composition program. For this high school group, the decision of interest for valid decision making was: How do the writing skills of students identified as competent in high school compare with the skills of the college writing groups to which they are expected to transfer? The decision of interest of college writing groups was: How are those students identified?

In order that a system can accurately classify examinees into the correct writing groups, it must be able to distinguish among the groups. It was found that three of the systems, General Impression, the Diederich Expository Scale and the CSE Analytic scale were associated with differences in group performance. T-unit was not, making it virtually useless for classification purposes. The scoring systems, when used to discriminate among groups, were effective in separating members of the High School Low group from all of the others. Their ability to separate clearly members of the High School High, College Low and College High groups, however, is less clear. The classification accuracy of three systems (GI, DES, CSE) is quite poor for College High and only fair for High School High and College Low.

There are two explanations for the lack of group separation among the High School High, College Low and College High examinees. One is that the members of the College High group were not truly representative of College High writers. These writers might have been reclassified more accurately as High School High or College Low. Results support this conclusion (see Tables 17 and 18). Most of the misclassifications of College High writers were into the College Low group. Two writers, however, were classified as High School Low. The presence of these two aberrant samples drastically affects the results of the discriminant and classification analyses in a relatively small sample.

A second explanation for the lack of group separation for the three groups is that members of both the College High and Low groups are chosen by the university placement Subject A examination from the pool of High School High writers. Thus members of the High School High group may possess writing skills identical to either those of the College High or Low writers rather than of lower quality. Thus, High School High writers may be indistinguishable from college writers due to the very nature of the writing skills possessed by examinees at that level.

When the scoring systems are used to classify examinees into the proper criterion groups, it was found that three systems classify more accurately than two. This finding was predictable in that classification accuracy is often enhanced by the use of several sources of information for the placement decision. Still another explanation for the advantage of using three scoring systems to classify examinees is tied to the kinds of scoring systems used in this study. Two analytic and a general impressionistic system were retained for the discriminant analyses. Although the analytic system (the Diederich Expository Scale) contributed most to group separation, the best prediction of group membership was obtained with a combination of

DES and General Impression scores. It may be that the accurate assessment of writing competence must incorporate both impressionistic and analytic strategies. Support for this interpretation is gained when it is remembered that the one system that appears as the strongest discriminating variable is the Diederich Expository Scale, a system that combines elements included in both the general impression and analytic scales. The CSE scale, while contributing to group separation, does not lead to as accurate classification of group membership as the DES. This result may be due, in part, to the fact that the scale provides no rating category for an overall impression of the effectiveness of the essay or judgment of writing style.

These interpretations must be accepted with caution, however, due to the unusual intercorrelations of the scoring systems. There are group by scoring system differences which may diminish the possibilities that any one scoring system can be an accurate predictor for all four groups. It does appear that either holistic or analytic methods are more effective for distinguishing among writers of several ability levels and within a restricted age range than are measures of syntactic maturity.

The results of the classification analysis indicate that it is fairly easy to identify the lowest ability writers in a group and even, perhaps, some of the highest, but most difficult to separate writers of medium to high ability. This assertion must be modified by the fact that the College High criterion group may not be a truly representative sample of high ability writers. High school grouping practices provide yet another reason for the differential classification accuracies obtained between the high school and college groups. In high school, students are assigned to English classes on the basis of previous grades, teacher recommendation, and post high school educational plans. The high school environment is smaller; teachers often know students well enough to "counsel" them into the appropriate English classes. On the other hand, the university composition placement decision is made on the basis of an impersonal, one-shot examination scored on criteria unknown to students and teachers. This procedure greatly increases the possibility for mistakes in classifying examinees.

The kinds of students found in high school "low" writing programs are very different from either High School High ability writers or college writers. These students do not always intend to attend college and so find themselves in classes where the emphasis is not upon expository writing. Additionally, high school remedial writing courses focus on the improvement of control over mechanics, grammar and organization--rudiments of writing mastered by students in pre-college composition courses. The career aspirations and value orientation of high school high students are much closer to those of the college groups than to their high school counterparts. Thus, the great separation between high school low writers and all of the other groups may be more a function of motivational and sampling differences than of scoring systems used to describe their performance. What is of real interest is the discrimination among the middle and high level writers belonging to the High School High, College Low and College High criterion groups.

Classification accuracy is best for the low writing groups in both high school and college. It may be that the characteristics of "low" writing performance are more stable and discernible than those of "high writers". To paraphrase Tolstoy, poor essays are very much alike but good essays can be outstanding in many different ways.

Recommendations

This study, like so many others, raises more questions than are answered. The dominant issue to emerge from the comparison of different approaches to the assessment of writing competence is what kind of decision must be made on the basis of writing assessment? This question assumed importance in the present study when it became apparent that current methods of assigning college students to remedial or regular composition classes do not adequately discriminate between the low and high level college writers when other systems are used. It must be questioned whether an examination designed to sort examinees, though reliable, can be valid if its discriminatory power is weak. Thus, future studies should address the validity issues related to the purposes of the testing program as well as those associated with the characteristics of the scoring systems themselves.

One encouraging result of the comparison of different approaches to scoring writing samples is that all systems can be highly reliable. These results support the finding of Follman and Anderson (1967) and the observations of Cooper (1977). Yet it is clear from the results of this study that, although almost any kind of scoring system can be highly reliable with rater training, reliability increases as scoring rubrics become more explicit. Thus, the differential reliabilities of scoring systems, while not appreciable, do exist. However, it is not the reliabilities, but the validities of the systems that are still very much at issue.

The ability of measures of syntactic maturity to assess writing quality is seriously questioned by the results of this study. The lack of correlation of T-unit scores with three widely used qualitative scoring systems indicates that claims for a relationship between writing quality and syntactic maturity, at least for writers within a restricted age range, must be seriously reexamined.

General Impression scoring, a widely used screening or grading procedure with ease of training and speed of scoring, may not be the best system for making graduation or placement decisions. The lack of descriptive power of general impression rubrics severely restricts the instructional utility of these systems when compared to analytic scoring techniques. Results of this study indicate, however, that, while not sufficient as a placement tool, GI scores may be a necessary part of any writing assessment procedure.

The Diederich Expository Scale, while requiring extensive training time, appears to be the best system for distinguishing between high and low performance writers at both the high school and college levels. Its mild correlations with General Impression and CSE analytic scores indicate that the DES may be providing two sorts of information, impressionistic as well as analytic. It was found that discrimination and classification accuracy were increased when the number of data sources were increased. These findings suggest that the complex construct "writing skill" might best be assessed by more than one approach in order to better define the domain of exposition.

The use of the CSE scale to distinguish among examinees at different writing skill levels is a highly reliable and promising technique. The fact that this scale contributes less to group discrimination than the Diederich Expository Scale is possibly explained by its lack of attention to any of the qualitative aspects of writing such as style, originality, or ability of the essay to create an impression of "good writing." A revision of the scale with a category for the global qualitative aspects measured by General Impression or the DES may increase the classification and descriptive powers of this system.

Suggestions for Further Research

Of primary importance in further research is the investigation of group by scoring system differences hinted at in the fluctuating intercorrelations and irregular score distributions reported in the previous section. A second line of investigation is the search for scoring systems that will discriminate among the High School High, College Low, and College High groups. Of special interest is, what combination of impressionistic and analytic dimensions best describes the writing performance of middle to high ability writers while at the same time retains the ability to make fine-grained discriminations among groups?

Finally, the instructional validities of the scoring systems require investigation. The scoring systems in the present study yield different kinds of information about student writing. The amount of guidance for instructional decision making provided by the systems also varies. The important question is, how much and what kinds of information are necessary for the planning and implementation of effective writing programs? For once accurate placement decisions are made, the goal of measurement is to provide information that will enable teachers to design optimal instructional sequences for each student. The descriptive and prescriptive validities of the present scoring systems are still at issue. Until these validities are documented, writing assessment programs used to make diagnostic-placement decisions will lack credibility, humanity, and integrity.

REFERENCES

Alpren, P. F. Can children be helped to increase the originality of their story writing? Research in the Teaching of English, #7 (1973), 372-86.

Anderson, C. C. The new STEP essay test as a measure of Composition ability, Educational and Psychological Measurement, 1960, 20, 95-102.

Bateman, D. R., and Zodonis, F. J. The effect of a knowledge of generative grammar upon the growth of language complexity. Columbus: The Ohio State University, U. S. Office of Education Cooperative Research Project, 1964.

Braddock, R. Evaluation of writing tests. In A. Grommon (Ed.) Review of selected published tests in English, Urbana, Illinois: National Council of Teachers of English, 1976.

Braddock, R., Lloyd-Jones, R., and Schoer, I. Research in written Composition. Urbana, Illinois: National Council of Teachers of English, 1963.

Brooks, C., and Warren, R. P. Modern rhetoric: Shorter edition. New York: Harcourt, Brace & World, 1961.

Coffman, W. B. Essay examinations. In R. L. Thorndike (Ed.), Educational measurement. (2nd. ed.) Washington, D. C.: American Council of Education, 1971.

Conlan, G. How the essay in the CEEB English test is scored. Princeton, New Jersey: Educational Testing Service, 1976.

Cooper, C. Holistic evaluation of writing. In C. Cooper and L. Odell (Eds.), Evaluating writing. State University of New York at Buffalo, 1977.

Cooper, C. Measuring growth in writing. English Journal, 1975, 64 (3).

Cooper, C. and Odell, L. (Eds.). Evaluating writing: Describing, measuring, judging. Buffalo: State University of New York at Buffalo, 1977.

Cronbach, L. Test validation. In R. L. Thorndike (Ed.). Educational measurement (2nd. ed.) Washington, D. C.: American Council on Education, 1971.

Diederich, P. B., French, J. W., and Carlton, S. Factors in judgments of writing ability. Princeton, New Jersey: Educational Testing Service, 1961.

Diederich, P. B. Measuring growth in English. Urbana, Illinois: National Council of Teachers of English, 1974.

Findlayson, D. S. The reliability of the marking of essays. British Journal of Educational Psychology, 1951, 21, 126-134.

Follman, J. C., and Anderson, J. A. An investigation of the reliability of five procedures for grading English themes. Research in the Teaching of English, 1967, 190-200.

French, J. W. Schools of thought in judging excellence of English themes. In Proceedings of the 1961 Invitational Conference on Testing Problems. Princeton, New Jersey: Educational Testing Service, 1962, 19-28.

Frogner, E. Problems of sentence structure in pupils' themes. English Journal, 1933, 22, 742-749.

Godshalk, F., Swineford, F., and Coffman, W. The Measurement of writing ability. New York: College Entrance Examination Board, 1966.

Gottesman, J., and Schilling, M. A Common Ground for Assessing Competence Written Expression. Los Angeles: Office of the Los Angeles County Superintendent of Schools, 1978.

Gorrell, R. Not by nature: Approaches to rhetoric. English Journal. 1966, LV, 409-16, 449.

Hambleton, R., Swaminathan, H., Algina, J., and Coulson, D. Criterion referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1977, 48 (1).

Harris. C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin, and W. J. Popham (Eds.). Problems in criterion-referenced testing. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Howerton, M. C., Jacobson, M., and elden, R. The relationship between quantative and qualitative measures of writing skill. Paper presented at the Annual Meeting of the American Educational Research Association, New York, April, 1977.

Hunt, K. W. Differences in grammatical structures written at three grade levels, the structure to be analyzed by transformational methods. Tallahassee, Florida State University, U. S. Office of Education Cooperative Research Project 1998, 1964.

Hunt, K. W. Grammatical structures written at three grade levels Research Report No. 3. Urbana, Illinois: National Council of Teachers of English, 1965.

Hunt, K. W. Early blooming and late blooming syntactic structures. In C. Cooper and L. Odell (Eds.), Evaluating writing. State University of New York at Buffalo, 1977.

Hunt, K. W., and O'Donnell, R. An elementary school curriculum to develop better writing skills. Tallahassees, Florida State University: U. S. Office of Education Grant #4-9-08-903-0042-010, 1970.

Kincaid, G. L. Some factors affecting variation in the quality of students' writing. In R. Braddock, R. Lloyd-Jones, and I. Schoer, Research in Written Composition. Urbana, Illinois: National Council of Teachers of English, 1963.

Linn, R. Issues of validity in measurement for competency-based programs. Paper presented at the Annual Meeting of the National Council of Measurement in Education, New York, April, 1977.

LLoyd-Jones, R. Primary trait scoring. In C. Cooper and L. Odell (Eds.) Evaluating Writing. State University of New York at Buffalo, 1977.

Mellon, J. C. Methodological problems in relating grammar instruction to writing ability: A critical review of selected research. Graduate School of Education, Harvard University, Cambridge, Massachusetts, 1965.

Meyers, A. E., McCanville, C. B. and Coffman, W. Simplex structure in the grading of essay tests. Educational and Psychological Measurement, 1966, 24, 41-54.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, California: McCurchan, 1974.

Nail, P., Fitch, R., and Halverson, J., et. al. A scale for the evaluation of high school student essays. Urbana, Illinois: National Council of Teachers of English, 1960.

O'Donnell, R.C., Griffin, W. J., and Norris, R. C. A systax of kindergarten and elementary school children. A transformational analysis. Research Report No. 8. Urbana, Illinois: National Council of Teachers of English, 1967.

O'Hare, F. Sentence combining: Improving student writing without formal grammar instruction. NCTE Research Report No. 15, Urbana, Illinois: National Council of Teachers of English, 1973.

Perry, D. K., and Cannon, W. M. Vocational interests of female computer programmers. Journal of Applied Psychology, 1968, 52, 31-35.

Popham, W. J. Educational evaluation. New Jersey: Prentice-Hall, 1975.

Sager, C. Improving the quality of written composition through pupil use of rating scale. Ann Arbor, Michigan: University Microfilms, Order #73-23-605, 1973.

Sanderson, G. Exposition and the English language. (2nd. ed.) New York: Appleton-Century-Crofts, 1969.

Seegers, J. C. Form of discourse and sentence structure. Elementary English Review, 1933, 10, 51-54.

Smith, L. S. Investigation of writing assessment strategies. Studies in Measurement and Methodology Work Unitl: Design and Use of Tests. NIE Contract #OB-NIEG-78-0213. Center for the Study of Evaluation, University of California, Los Angeles, November, 1978.

Smith, R. Grading the advanced placement English examination. Princeton, New Jersey: College Board Publications, 1975.

Troxler, A., and Anderson, H. The reliability of an essay test in English. School Review, 1935, XLIII, 534-39.

Veal, L. R. and Tillman, M. Mode of discourse variation in the evaluation of children's writing. Research in the Teaching of English, 1971, 5, 37-45.

Warriner, J., and Griffith, F. English grammer and composition. New York: Harcourt-Brace-Janovich, 1973.

Wilson, H. A. A judgmental approach to criterion referenced testing. In C. W. Harris, M. C. Alkin, and W. J. Popham (Eds.). Problems in criterion-referenced testing. CSE Monograph Series in Evaluation No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Wiseman, S., and Wrigler, J. Essay reliability: The effect of choice of essay title. Educational and Psychological Measurement, 1958, XVIV, 129-188.

APPENDIX A

Directions to Students and Writing Prompts

Your Name: ______________________________

Course Title: ____________________________

Date: ___________________________________

Professor: ______________________________ Your School: ____________

DIRECTIONS

1. Write an essay on the topic given on the next page.
2. Your essay should be at least one page long (about 200 words).
3. Write clearly in pen on the paper provided.
4. You will have 45 minutes to write your essay.

TOPIC 1

Reading, writing, and arithmetic, the three R's, have become a serious problem for many young adults. In fact, every year thousands of American students graduate from high school unable to read simple words, write short essays, or do math. To understand this problem, experts have studied areas such as teaching methods, family life, and the effects of television. However, no one best explanation has been found.

Assume you have been asked by a group of parents and teachers to explain the problem. Write an essay in which you explain why many of today's students graduate from high school without basic skills. Your essay should focus on one (or more) of the "three R's", reading, writing, or math.

Remember, the purpose of your essay is to give an informative explanation, not opinions. Back up your ideas with specific support, such as examples, facts, and other details.

TOPIC 2

The use of alcohol, drugs, and cigarettes, has become a serious problem for many young adults. In fact, every year the use of these substances hurts the lives of thousands of American teenagers by making them emotionally and physically ill. To better understand this problem, experts have studied areas such as family and school life, pressure from friends, and the effects of television and advertising. However, no one best explanation has been found.

Assume you have been asked by a group of parents and teachers to explain the problem. Write an essay to help explain why many of today's teenagers abuse alcohol and drugs or smoke cigarettes. Your essay should focus on one (or more) of the substances, alcohol, drugs, or cigarettes.

Remember, the purpose of your essay is to give an informative explanation, not opinions. Back up your ideas with specific support, such as examples, facts, or other details.

APPENDIX B

Agenda for Rater Training

AGENDA FOR RATER TRAINING SESSIONS

1. Read the rubric provided for the scoring system and discuss the meanings of each of the ratings.
2. Read writing prompts and discuss what characteristics a model essay fitting each level of the scoring system would possess.
3. Read sample papers and using scoring system to assign a rating.
4. Compare ratings.
5. Discuss reasons for differences among ratings.
6. Read another sample paper, score, discuss scores.
7. Continue procedure until readers agree that they understand the meanings of the rating categories and they agree among themselves to assign scores in the same manner.
8. Lunch Break.
9. Pilot test for obtaining reliability estimates of raters on the scoring system. Read 24 papers without discussion and rate.

APPENDIX C

Dependent Measures

1. General Impression Rating Scale and Amended Rubric
2. ETS Composition Scale (Diederich Expository Scale) and Abbreviated Rubric
3. CSE Analytic Scale and Abbreviated Rubric
4. T-Unit Analysis Procedures

Impressionistic Rating Scale

Rater Materials

Impressionistic Rating Procedures

The purpose of Impressionistic Rating procedures is to form a single impression of a piece of writing as to how well it communicates a whole message to the reader. Impressionistic scoring assumes that each characteristic that makes up an essay -- organization of ideas, content, mechanics and so on -- is related to all other characteristics. Impressionistic scoring further assumes that the qualities of an essay cannot easily be separated from each other. In short, the procedure views a piece of writing as a total work, the whole of which is greater than the sum of its parts.

Discerning readers naturally will attend to, or be influenced by, some essay characteristics more than others. In Impressionistic scoring, however, readers are not expected to assign values to separate characteristics of an essay. Instead, they should arrive at a judgment regarding the overall quality of the essay.

For this rating session, you are being asked to form an overall impression concerning the effectiveness of the essays as examples of expository writing.

The Topics

You will be reading essays on two different topics. Both topics, though, were designed to elicit writing in the expository mode. Some views on exposition are given below.

.Exposition is the kind of discourse that explains or clarifies a subject.

.Exposition seeks to explain or inform through such methods as giving reasons or examples, comparing and contrasting, defining, ennumerating or through a combination of methods.

.Exposition explains why or how.

.Exposition promotes reader understanding of a subject.

Instructions to Readers

You are to read each essay quickly in order to form an overall impression of its quality. To assign the essay a score, consider the following question: To what extent is the essay an example of effective exposition?

Assign each paper a mark of 1 - 6 using the scale below.

High

6 = An excellent example of exposition

5 = A good example of exposition

Middle

4 = An adequate example of exposition

3 = A minimally adequate example of exposition

Low

2 = A poor example of exposition

1 = Barely readable and/or off the topic

PLEASE NOTE: The papers may not represent a "normal" distribution, that is, some categories may have very few papers while others may have many. In other words, do not intentionally sort an equal number of papers into all six categories.

AMENDED IMPRESSIONISTIC RUBRIC

The following descriptors were used in addition to the rubric provided by the experimenter. Original guidelines are underlined.

Assign each paper a mark of 1 - 6 using the scale below:

High

6 = An excellent example of exposition

It has style, is well organized and a solidified main idea.

5 = A good example of exposition

It has some flair. The stuff is there but it may need some rearranging. It's smooth, and eoesn't disgrace the teacher. This student doesn't need supervision when writing. Has all the components of the essay but may be badly arranged.

Middle

4 = An adequate example of exposition

Elements of the essay are there but they need expansion. Essay is workmanlike but may have some stupid errors. Has at least two well developed points and doesn't discredit the teacher.

3 = A minimally adequate example of exposition

Essay show no control of the language. It has structural problems.

Low

2 = A poor example of exposition

Essay is short, perhaps one paragraph. Ideas are dashed onto the page. It is not incoherent.

1 = Barely readable and/or off the topic

Essay is too personal, short and has many spelling errors. Essay uses street English as opposed to Standard Written English.

ETS COMPOSITION SCALE

Topic ___________ Reader's initials _________ Paper ___________

GENERAL MERIT	Low		Middle		High	
Ideas	2	4	6	8	10	-
Organization	2	4	6	8	10	-
Wording	1	2	3	4	5	-
Flavor	1	2	3	4	5	-
MECHANICS						
Usage-Sentences	1	2	3	4	5	-
Punctuation	1	2	3	4	5	-
Spelling	1	2	3	4	5	
					Total	

GENERAL MERIT

1. Ideas

High

The student has given some thought to the topic and has written what he really thinks.

He discusses each main point long enough to show clearly what he means.

He supports each main point with arguments, examples, or details; he gives the reader some reason for believing it.

His points are clearly related to the topic or to the main idea or impression he is trying to get across.

No necessary points are overlooked and there is no padding.

Middle

The paper gives the impression that the student does not really believe what he is writing or does not fully understand what it means.

He tries to guess what the teacher wants and writes what he thinks will get by.

He does not explain his points very clearly or make them come alive to the reader.

He writes what he thinks will sound good, not what he believes or knows.

Low

It is either hard to tell what points the student is trying to make or else they are so silly that he would have realized that they made no sense if he had only stopped to think.

He is only trying to get something down on paper.

He does not explain his points; he only writes them then goes on to something else, or he repeats them in slightly different words.

He does not bother to check his facts, and much of what he writes is obviously untrue.

No one believes this sort of writing -- not even the student who wrote it.

2. Organization

High The paper starts out at a good point, has a sense of movement, gets somewhere, then stops.

The paper has a plan that the reader can follow; he is never in doubt as to where he is or where he is going.

Sometimes there is a little twist near the end that makes the paper come out in a way that the reader does not expect, but it seems quite logical.

Main points are treated at greatest length or with greatest emphasis, others in proportion to their importance.

Middle The organization of this paper is standard and conventional.

There is usually a one-paragraph introduction, three main points, each treated in one paragraph, and a conclusion that often seems tacked on or forced.

Some individual points may be treated in greater detail than important points and there is usually some dead wood that might be better cut out.

Low This paper starts anywhere and never gets anywhere.

The main points are not clearly separated from one another and they come in a random order -- as though the student had not given any thought about what he intended to say before he started to write.

The paper seems to start in one direction, then another, then another until the reader is lost.

3. Wording

High

The writer uses a sprinkling of uncommon words or common words in an uncommon setting.

He shows an interest in words and in putting them together in slightly unusual ways.

Some of his experiments with words may not quite come off, but this is such a promising trait in a young writer that a few such mistakes may be forgiven.

For the most part he uses words correctly, but he also uses them with imagination.

Middle

The writer is addicted to tired old phrases and hackneyed expressions.

If you left a blank in one of his sentences almost anyone could guess what word he would use at that point.

He does not stop to think how to say something; he just says it in the same way as everyone else.

A writer may also get a middle rating on this quality if he overdoes his experiments with uncommon words: if he always uses a big word when a little word would serve his purpose better.

Low

The writer uses words so carelessly or inexactly that he gets far too many wrong.

These are not intentional experiments with words in which failure may be forgiven; they represent groping for words and using them without regard to their fitness.

A paper written entirely in a childish vocabulary may also get a Low rating, even if no word is clearly wrong.

4. Flavor

High	The writing sounds like a person, not a committee. The writer seems quite sincere and candid, and he writes about something he knows -- often from personal experience. You could not mistake this writing for the writing of anyone else. Although the writer may play different roles in different papers, he does not put on airs. He is brave enough to reveal himself just as he is.
Middle	The writer usually tries to appear better or wiser than he really is. He tends to write lofty sentiments and broad generalities. He does not put in the little homely details that show he knows what he is talking about. His writing tries to sound impressive. Sometimes it is impersonal and correct but colorless, without personal feeling or imagination.
Low	The writer reveals himself well enough but without meaning to. His thoughts and feelings are those of an uneducated person who does not realize how bad they sound. His way of expressing himself differs from standard English, but it is not his personal style; it is the way uneducated people talk in his neighborhood. Sometimes the unconscious revelation is so touching that we are tempted to rate it high on flavor, but it deserves a high rating only if the effect is intentional.

MECHANICS

5. Usage, Sentence Structure

High There are no vulgar or illiterate errors in usage by present standards or informal, written English, and there are very few errors in points that have been emphasized in class.

The sentence structure is usually correct, even in varied and complicated sentence patterns.

Middle There are a few serious errors in usage and several in points that have been emphasized in class but not enough to obscure meaning.

The sentence structure is usually correct in the more familiar sentence patterns, but there are occasional errors in more complicated patterns, as in parallelism, subordination, consistency of tenses, reference of pronouns, etc.

Low There are so many serious errors in usage and sentence structure that the paper is hard to understand.

6. Punctuation, Capitals, Abbreviation, Numbers

High There are no serious violations of rules that have been taught -- except slips of the pen.

Note, however, that modern editors do not require commas after short introductory clauses, around nonrestrictive clauses, or between short coordinate clauses unless their omission leads to ambiguity or makes the sentence hard to read.

Contractions are acceptable, often desirable.

Middle There are several violations or rules that have been taught -- as many as usually occur in the average paper.

Low Basic punctuation is omitted or haphazard, resulting in fragments, run ons, etc.

7. Spelling

High	The high paper usually has not more than five misspellings and these are confined to words that are hard to spell. The spelling is consistent; words are not spelled correctly in one sentence and misspelled in another --unless the misspelling appears to be a slip of the pen. If a poor paper has not misspellings it gets a high rating on spelling.
Middle	There are several spelling errors in hard words and a few violations of basic spelling rules but no more than one finds on the average paper.
Low	There are so many spelling errors they interfere with comprehension.

Abbreviated Rubric For ETS Scale

IDEAS

2. Thoughts without reasons
   Rambling
   No clear separation of ideas
   Unloading
   Doesn't develop or move on from main point
   Presents only one idea and its poorly supported
4. Has at least 2 good ideas but with less than average support
6. Has at least 3 points with at least one supportive statement for each point.
   Must have a thesis statement
8. Must have a thesis statement
   Has to have 3 points
   Has support for at least 2 statements which consists of a topic sentence with a paragraph attached.
10. Concrete evidence is provided
    Good support is provided for all 3 or more points in the essay

ORGANIZATION

2. Brainstorming
   A one-paragraph essay that is not well-developed, although several points may be mentioned
   Essay has no introduction and/or conclusion
4. There is an introduction, 3 points, conclusion in one paragraph if the essay consists of one paragraph

   Has at least one successful paragraph

   Has beginning and end, but may lack thesis statement
6. Must have thesis statement
   Has 3 points and a conclusion
   May be mechanically arranged, abrupt transitions
8. Gets the job done in 50 minutes
   Introduction may be mechanical
   Uses transitions.

FLAVOR

1. Flat, colorless, non-standard, street-talk
2. Common, but not distinctive, writing
   No distinguishing traits or features
3. At least one sentence is expressed originally and has flair
4. Smooth, has flow, can be a female reader in the positive sense, has character
5. Polished and distinctive

WORDING

1. Street talk, oral style
2. Awkward, preposition problems, diction problems
3. Heavy reliance and predictible language
4. Uses words correctly and with imagination
   Accurate, precise
5. Interesting words, put together in unusual ways

USAGE

Look at: misuse of commas, fragments, misplaced modifiers, agreement, reference parallelism

1. Street talk
2. Glaring spots of poorly-written sentences
3. May be awkward but attempts difficult sentence constructions

PUNCTUATION

1. Ignorance of conventions
2. May have punctuation mistakes but compensating factors such as use of correct punctuation in difficult constructions, etc.
   Has about 5 errors

SPELLING

1. Illiterate
2. Too many errors
3. Spelling mistakes but compensates with other difficult but correct words
4. Spelling errors confined to fairly difficult words and are consistent
5. 5 or less mistakes on hard words, for 1/2 page essays, 3 or less mistakes

THE ANALYTIC RATING SCALE

Analytic Scale Elements

1. Essay Focus: The introduction or conclusion of the essay clearly indicates the subject and main idea of the whole essay.

2. Essay Development: All major subtopics ("main points") clearly relate to the main idea of the whole essay.

3. Essay Organization: The main idea is developed according to a clearly discernible method of organization.

4. Support: Generalizations and assertions are supported by specific clear supporting statements.

5. Paragraphing: The essay is composed of one or more clearly discernible units of thought, e.g., paragraphs.

6. Mechanics: The essay is free of intrusive mechanical errors.

COMPLETE VERSION OF ANALYTIC SCALE

ELEMENT 1

Essay Focus

The introduction (if deductively structured) or conclusion (if inductively structured) of the essay clearly indicates the subject and main idea of the whole essay.

4. The introduction (and/or conclusion) of this paper clearly conveys the main idea of the whole essay. It also limits the topic by alerting the reader to the key points covered in the body of the essay.

   Specifically, in the introduction (and/or conclusion):

   a. The subject of the essay is clearly identified.

   b. The main idea of the whole essay is clearly stated or implied.

   c. The topic is clearly limited. That is, key points (e.g., reasons, ideas) or major line(s) of reasoning treated in the essay are identified or summarized.

3. The introduction (and/or conclusion) of this paper conveys the main idea of the whole essay. It sets limits on the topic, but does not clearly suggest how the main idea is developed.

   Specifically, in the introduction (and/or conclusion):

   a. The subject of the essay is clearly identified.

   b. The main idea of the whole essay is clearly stated or implied.

   c. An attempt is made to limit the topic. That is, the number -- or type -- of key points is specified, but there is not clear reference to the substantive issues treated in the body of the essay.

2. The introduction (and/or conclusion) of this paper gives the reader a fairly clear sense of the main idea of the whole essay. However, neither the introduction nor the conclusion help focus -- or bring direction to -- the body of the paper.

   Specifically, in the introduction (and/or conclusion):

   a. The subject of the essay is identified.

   b. The main idea of the whole essay is stated or implied.

   c. No attempt is made to limit the topic.

ELEMENT 1 (continued)

Essay Focus

1. Neither the introduction nor the conclusion is helpful to the reader in obtaining any sense of the main idea of the essay.

   Specifically, in the introduction (and/or conclusion):

   a. The subject of the essay is not clearly identified or there is no reference to the subject

   AND/OR

   b. The main idea of the whole essay is not clearly stated or implied or no reference is made to the main idea, or the reference is confusing.

ELEMENT 2

Essay Development

All major subtopics ("main points") clearly relate to the main idea of the whole essay.

4. All ideas and details in this paper are clearly and logically related to a single dominant "main idea." Development of the idea is not interrupted by digressions of thought or extraneous material.

   Specifically:

   a. All major subtopics ("main points") clearly relate to the main idea of the whole essay.

   b. The way in which subtopics contribute to development of the main idea, e.g., as illustration, as examples, as background, is readily evident.

   c. There are no digressions of thought or extraneous material.

3. All major ideas in this paper are related to the dominant "main idea." There may be a few minor digressions of thought, such as extended listings of supporting detail. However, digressions do not stray entirely away from the main point; they simply interrupt the logical flow of thought. One or two tangential ideas may be introduced but are not developed.

   Specifically:

   a. All major subtopics ("main points") clearly relate to the main idea of the whole essay.

   b. The way in which subtopics contribute to development of the main idea, e.g., as illustration, as examples, as background, can be easily inferred.

   c. There are no major digressions of thought, i.e., an idea tangential to the main idea of the whole essay which is developed at some length.

2. The relationship of many ideas and details to the overall main idea of the essay is somewhat tenuous. This may be due to one (or more) of the following types of problems. The paper may begin to develop a main idea, but "stray" or "wander" off the topic, failing to return to the main point. Or, the paper may be peppered with many tangential ideas/details which disrupt the development of subtopics (and, in turn, the main idea).

   Specifically:

   a. Some major subtopics ("main points") relate to the main idea of the whole essay.

   b. The way in which most subtopics contribute to development of the main idea is not readily evident.

Essay Development

c. Though parts of the paper help develop a single main idea, the essay may contain one or more major digressions of thought.

1. Few ideas or details in this paper contribute to development of a single main idea. In fact, the reader is likely to have difficulty discerning the main idea of the essay, if one is present. A paper which does not address the topic posed in the assignment also should receive a "1" rating.

Specifically:

a. There are few or no subtopics which relate to a dominant main idea AND/OR

b. A majority of the essay is devoted to ideas/details unrelated to the topic of the assignment.

ELEMENT 3

Essay Organization

The main idea is developed according to a clearly discernible method of organization.

4. The plan by which the essay is structured is readily apparent and consistently applied throughout the essay. Some logical scheme serves as the underlying basis for the overall essay structure. That is, there is an easily discernible logical flow from one idea (subtopic) to the next.

   Specifically:

   a. The plan for organizing the essay is clearly established in the beginning, e.g., through reference to the subtopics or line(s) of reasoning to be developed

      AND/OR

      The plan for organizing the essay is readily evident to the reader due to the effective use of transitions and linking expressions.

   b. The plan by which the essay is organized is consistently applied throughout the essay.

   c. The overall structure of the essay reflects some logical, underlying organizing rubric. That is, subtopics are presented according to some logical scheme, or order, such as: order of importance, seriousness; comparison and contrast; from specific to general; in chronological order; through classification.

3. The essay is organized according to some recognizable plan which is employed throughout the essay. Some organizing rubric is used to divide the content/ideas of the essay into units of thought. For example, each subtopic might represent a different reason or category of content. Subtopics, however, are not presented in any particular logical order.

   Specifically:

   a. The plan for organizing the essay is established in the beginning, e.g., through reference to the subtopic(s) or line(s) of reasoning to be developed

      AND/OR

      The plan for organizing the essay is evident to the reader due to the use of transitions and linking expressions.

   b. The plan by which the essay is organized is employed throught the essay.

   c. Major subtopics are not presented according to any underlying logical order; they could be "reshuffled" without dramatically altering the overall impact or clarity of the essay.

Essay Organization

2. The essay is not structured according to any readily discernible plan. However, the reader is able to infer some association between major ideas. The logical flow from one idea to the next is so tenuous that the reader must supply his/her own links or transitions to form a clear relationship among ideas.

   Specifically:

   a. A plan for ordering ideas/details is attempted; the plan is not readily apparent to the reader. The essay does not make effective use of transitions or linking expressions to help the reader follow the trend of thought
   AND/OR

   b. The plan for ordering ideas/details is not consistently applied throughout the essay.

1. No plan for organizing ideas and details is apparent. Subtopics are not arranged in any discernible order. The reader has difficulty inferring any relationship among ideas or any pattern of thought.

   Specifically:

   a. No plan for organizing ideas and details is evident. Transitions and linking expressions, if present, are not helpful in following the writer's train of thought.

ELEMENT 4

Support

Generalizations and assertions are supported by specific, clear supporting statements.

4. All generalizations and assertions in this paper have logically related support. Supporting statements clearly are at a greater level of specificity than the generalizations they are intended to support. The writer makes effective use of concrete detail to support ideas.

   Specifically:

   a. Generalizations and assertions are supported by more specific statements.

   b. Supporting statements provide specific detail, such as examples, facts, anecdotes, and/or employ concrete language to convey ideas.

3. Most generalizations and assertions are supported by more specific, clear and logically related statements.

   Specifically:

   a. Generalizations and assertions are supported by more specific statements.

   b. Many supporting statements provide specific detail, such as examples, facts, ancecdotes, and/or employ concrete language to convey ideas.

2. Support for many generalizations and assertions in this paper is not especially convincing. Supporting statements lack precision, either in language or in the use of specific detail.

   Specifically:

   a. An attempt is made to support generalizations and assertions through the use of statements at a greater level of specificity than the generalization they are intended to support.

   b. Supporting statements lack precision and clarity. The use of specific detail, such as examples, facts, or anecdotes, may be lacking and/or the language may be vague.

1. Support is provided for few or none of the generalizations or assertions. Few (or none) of the supporting statements are at a greater level of specificity than the generalizations themselves.

   Specifically:

   a. Supporting statements are provided for few or none of the generalizations or assertions.

   b. Supporting details, if present, are vague, confusing, or not logically related to the generalizations they apparently are intended to support, and/or the language of supporting statements is imprecise.

ELEMENT 5

Paragraphing

The essay is composed of one or more clearly discernible units of thought, e.g., paragraphs.

4. This paper contains one or more clearly circumscribed subtopics which are developed by cohesive groups of supporting statements. Each subtopic represents a complete unit of thought. Major units of thought are delineated by conventions of paragraphs.

   Specifically:

   a. All statements directly related to the same idea (subtopic) are contiguous, i.e., are presented together as logically related units or blocks, of thought.

   b. Conventions of papragraphs are used to differentiate major units of thought.

3. This paper is composed of one or more subtopics which are developed by supporting statements. Each subtopic represents a complete unit of thought. Distinctions between major units of thought would be clearer to the reader if paragraphs were more effectively employed.

   Specificially:

   a. All statements directly related to the same idea (subtopic) are contiguous, i.e., are presented together as logically related units, or blocks, of thought.

   b. There are instances where a paragraph is needed (i.e., to separate distinct units of thought: but is absent, and/or where a paragraph is present, but not warranted.

2. This paper contains one or more subtopics which represent major units of thought. However, the distinction between different units of thought is not consistently clear. Distinctions may be blurred because topic sentences and related support are not presented as cohesive and logically related groups of sentences.

   Specifically:

   a. Statements directly related to the same idea (subtopic) are not consistently presented together as logically related units, or blocks, of thought. For example, the essay might present a series of "topic sentences," which then are followed by supporting statements. The statements should be re-arranged, matching topic sentences to their own intended supporting statements.

   b. There are instances where a paragraph is needed (i.e., to separate distinct units of thought), but is absent, and/or where a paragraph is present but not warranted.

Paragraphing

1. The essay does not appear to contain any clearly differentiated units of thought, e.g., subtopics and related support. Statements appear to be presented in random order, with no clear relationship between topic sentences and supporting statements.

   Specifically:

   a. The essay contains few or no clearly differentiated units of thought. That is, sentences related to the same major idea are scattered throughout the essay.

   b. Conventions of paragraphs are absent or consistently incorrectly employed.

ELEMENT 6

Mechanics

The essay is free of intrusive mechanical errors.

4. The writer appears to have control over the mechanical aspects of this essay.

   Specifically:

   a. There are only a few minor errors in mechanics or no errors at all.

   b. Errors, if present, do not interfere with the clarity of communication.

3. Mechanics are not a problem in this paper. However, the writer appears to be occasionally careless and makes a few minor errors which are readily evident to the discerning reader. The paper also may contain a few common errors in mechanics.

   Specifically:

   a. There are only a few obvious mechanical errors.

   b. Errors do not interfere with the clarity of communication, e.g., occasional confusion of subject-object pronouns ("between you and I..."), spelling errors on difficult words, misuse of colons and semicolons.

   c. There are no major errors, such as run-on sentences, inappropriate fragments, lack of subject-verb agreement on simple sentences.

2. This essay is flawed by errors in mechanics, although the paper does not strike the discerning reader as being illiterate.

   Specifically:

   a. There are obvious errors in mechanics throughout much of the essay.

   b. Errors occasionally detract from the clarity of the communication, such as confusing antecedents, consistent omission of key words.

   c. Many of the errors are major, e.g., many run-on sentences or inappropriate fragments, lack of subject-verb agreement on simple sentences.

1. Mechanical errors make this paper very difficult to read and understand.

   Specifically:

   a. There are excessive and obvious errors in mechanics throughout the essay; nearly every sentence contains some type of error.

   b. Errors consistently interfere with the writer's attempt to communicate.

   c. Errors are not restricted to one type of problem, such as run-on sentences. Errors are diverse in nature and major.

T-Unit Analysis

I. Procedure

- A. Count the number of words in the composition and record on the rating sheet.
- B. With a pen or pencil, section off all T-units.
- C. Count the number of T-units in the composition.
- D. Divide the number of words by the number of T-units to obtain the words-per-T-unit score.

II. Segmenting Rules

- A. Rules for counting words in the composition
  1. Compound nouns written as one word count as one word.
  2. Compound nouns written as two words and hyphenated words count as two words.
  3. Phrasal proper names count as one word.
  4. Dates, such as June 21, July 2, count as two words.
  5. Contractions such as he'd, shouldn't, count as two words.
  6. Garbles (unintellible graphemes) should not be counted.
- B. Rules for identifying T-units
  1. T-units are minimally terminable communication units: i.e. a single independent predication together with any aubordinate clauses grammatically related to it.
  2. One main clause plus any subordinate or non-clausal structure attached to or embedded in it counts as one T-unit.
  3. Simple or complex sentences count as one T-unit.
  4. Compound sentences count as two T-units.
  5. When two independent clauses are linked by a coordinating conjunction, count the coordinating conjunction as the first element of the second clause.
  6. If a fragment can be made into a clause with the addition of one word, add the word and count the fragment as one T-unit.
  7. Discard fragments that cannot be made into independent clauses with the addition of only one word.
  8. Discard garbles, i.e., unintelligible strings of words.
  9. Ignore punctuation.
  10. For direct discourse, the first expression after "he said" is counted as the direct object of 'he' and is considered as one T-unit. Marsha said, "I really like you John. However, Clarance's father is a millionaire and I like the idea of Palm Beach (3 units).

Appendix D

Results of Pilot Tests for Rater Reliability

1. Analysis of Variance Table for Topic Effects on General Impression Scoring
2. Reliability Coefficient Alpha for General Impression Scoring
3. Reliability Coefficient Alpha for Diederich Expository Scale
4. Reliability Coefficient Alpha for CSE Analytic Scale
5. Reliability Coefficient Alpha for T-Unit Analysis

Appendix D1

Analysis of Variance for Dependent Variable: General Impression

Source	SS	DF	F	Estimates of Variance Components
Subjects	98.614	11		0.980
Topics	.510	1		0.10
Raters	2.698	3		0.316
*S x T	12.114	11	4.84*	0.214
S x R	8.177	33	1.08	0.10
T x R	.365	3	0.53	0.88
STR	7.5	33		.228
(Error term for interactions)				

Program: BMD 08V

Appendix D2

Analysis of Variance Table for Alpha Coefficient: General Impression

Source	SS	DF	F	P
Between S's	49.307	11		
Within S's	5.437	36		
Between Measures (Raters)	1.349	3	3.6*	.02
Residual	4.048	33		

Alpha=.97 Standardized Item Alpha=.97

(SPSS Batch Release 7.8 Update Manual, March, 1977 Program: Reliability)

Appendix D3

Analysis of Variance Table for Alpha Coefficient: DES

Source	SS	DF	F	P
Between S's	1235.934	11		
Within S's	3900.938	36		
Between Measures (Raters)	2431.432	3	-18.0	.00
Residual	1469.505	33		

Alpha=.60 Standardized Item Alpha=.82

Appendix D4

Analysis of Variance Table for Alpha Coefficient: CSE Analytic

Source	SS	DF	F	P
Between S's	301.932	11		
Within S's	45.438	36		
Between Measures (Raters)	3.598	3	.94	.43
Residual	41.838			

Alpha=.95 Standardized Item Alpha=.96

Appendix D5

Analysis of Variance Table for Alpha Coefficient

Source	SS	DF	F	P
Between S's	142.334	11		
Within S's	16.487	36		
Between Measures (Raters)	6.009	3	6.3	.001
Residual	10.478	33		

Alpha=.98 Standardized Item Alpha=.98